Getting the most out of RNA-seq data analysis

Tsung Fei Khang, Ching Yee Lau

Background: A common research goal in transcriptome projects is to find genes that are differentially expressed in different phenotype classes. Biologists might wish to validate such gene candidates experimentally or use them for downstream systems biology analysis. Producing a coherent differential expression analysis from RNA-seq count data requires an understanding of how numerous sources of variation such as the replicate size, the hypothesized biological effect size, and the specific method for making differential expression calls interact. We believe an explicit demonstration of such interactions in real RNA-seq data sets is of practical interest to the biologist. **Results:** Using two large public RNA-seq data sets - one representing strong, and another mild, biological effect size, we simulated different replicate size scenarios and tested the performance of several commonly-used methods for calling differentially expressed genes in each of them. Our results suggest that if the biological effect size is expected to be mild, then RNA-seq experiments should focus on validation of differentially expressed gene candidates. At least triplicates must be used, and the differentially expressed genes should be called using methods with high positive predictive value (PPV) such as NOISeq or GFOLD. In contrast, when biological effect size was strong, differentially expressed genes mined from unreplicated experiments using NOISeq, ASC and GFOLD had between 30 to 50% mean PPV, an increase of more than 30-fold compared to the case of mild biological effect size. Among methods with good PPV performance, having triplicates or more substantially improved mean PPV to over 90% for GFOLD, 60% for DESeq2, 50% for NOISeq, and 30% for edgeR. At replicate size of six, We found DESeq2 and edgeR to be reasonable methods for calling differentially expressed genes at systems level analysis as their PPV and sensitivity trade-off were superior to the other methods'. **Conclusion:** When biological effect size is strong, NOISeq and GFOLD are effective tools for detecting differentially expressed genes in unreplicated RNA-seq experiments for validation work. Having triplicates or more enables DESeq2 and edgeR to detect sufficiently large numbers of reliable gene candidates for downstream systems level analysis. When biological effect size is weak, systems level investigation is not possible, and no meaningful result can be obtained in unreplicated experiments. Nonetheless, NOISeq or GFOLD may yield limited numbers of candidates with good validation potential when triplicates or more are

available.

Getting the most out of RNA-seq data analysis

Tsung Fei Khang[1*] **and Ching Yee Lau**[2]

[1]**Institute of Mathematical Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia.**
[2]**Institute of Biological Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia.**
[*]**Corresponding author: tfkhang@um.edu.my; Tel: +60379674171**

ABSTRACT

Background: A common research goal in transcriptome projects is to find genes that are differentially expressed in different phenotype classes. Biologists might wish to validate such gene candidates experimentally or use them for downstream systems biology analysis. Producing a coherent differential expression analysis from RNA-seq count data requires an understanding of how numerous sources of variation such as the replicate size, the hypothesized biological effect size, and the specific method for making differential expression calls interact. We believe an explicit demonstration of such interactions in real RNA-seq data sets is of practical interest to the biologist.
Results: Using two large public RNA-seq data sets - one representing strong, and another mild, biological effect size, we simulated different replicate size scenarios and tested the performance of several commonly-used methods for calling differentially expressed genes in each of them. Our results suggest that if the biological effect size is expected to be mild, then RNA-seq experiments should focus on validation of differentially expressed gene candidates. At least triplicates must be used, and the differentially expressed genes should be called using methods with high positive predictive value (PPV) such as NOISeq or GFOLD. In contrast, when biological effect size was strong, differentially expressed genes mined from unreplicated experiments using NOISeq, ASC and GFOLD had between 30 to 50% mean PPV, an increase of more than 30-fold compared to the case of mild biological effect size. Among methods with good PPV performance, having triplicates or more substantially improved mean PPV to over 90% for GFOLD, 60% for DESeq2, 50% for NOISeq, and 30% for edgeR. At replicate size of six, We found DESeq2 and edgeR to be reasonable methods for calling differentially expressed genes at systems level analysis as their PPV and sensitivity trade-off were superior to the other methods'.
Conclusion: When biological effect size is strong, NOISeq and GFOLD are effective tools for detecting differentially expressed genes in unreplicated RNA-seq experiments for validation work. Having triplicates or more enables DESeq2 and edgeR to detect sufficiently large numbers of reliable gene candidates for downstream systems level analysis. When biological effect size is weak, systems level investigation is not possible, and no meaningful result can be obtained in unreplicated experiments. Nonetheless, NOISeq or GFOLD may yield limited numbers of candidates with good validation potential when triplicates or more are available.

Keywords: biological effect size, biological replicate size, differential gene expression analysis, RNA-seq

INTRODUCTION

Elucidating key genes associated with variation between different biological states at the genomic level typically begins with the mining of high dimensional gene expression data for differentially expressed genes (DEG). For a long time, biologists have been using microarrays for gene expression studies, and over the years, the collective experience of the community has congealed into a set of best practices for mining microarray data (Allison et al., 2006). Hence, to determine optimal replicate size, one may use the SAM package (Tibshirani, 2006); to call DEG, the moderated t-test (Smyth, 2005, 2004) would be applied (Jeanmougin et al., 2014), producing p-values for each gene that adjust for multiple comparisons (Dudoit et al., 2014). When jointly considered with fold change (Xiao et al., 2014), the researcher can then get a set of DEG with strong potential to be validated by qPCR. Riding on such confidence, the

researcher could further study functional enrichment to gain understanding of dysregulated biological processes, or generate network-based hypotheses for targeted intervention.

Despite microarray's analytical maturity, RNA-seq - which is based on next-generation sequencing technology, is set to become the method of choice for current and future gene expression studies (Wang et al., 2009). In RNA-seq, direct transcript counting through mapping of short reads to the genome overcomes the problem of limited dynamic range caused by signal saturation in microarrays. In addition, the transcriptome can now be sequenced to unprecedented coverage, thus removing dependence on prior transcriptome knowledge which is crucial for probe design in microarrays. With the availability of numerous de novo transcriptome assembly tools (Li et al., 2014), meaningful gene expression studies in non-model organisms can now be done. While conceptually simple, sophisticated algorithms are involved in transforming raw reads to the final gene counts, and they constitute an important source of non-biological variation that must be appropriately accounted for (Oshlack et al., 2010).

Limited availability of biological material and costs of data production and bioinformatic support mean that RNA-seq data sets with little or no replication remain quite common today. Like microarray experiments, RNA-seq experiments that have less biological replicates are considered to have weak power for detecting genes with modest or weaker biological effect size. In fact, the problem may become worse from a multiple comparison point of view, as potentially many more genes are scored. Studies that aim at a systems level understanding using the list of DEG must therefore prioritize large replicate sizes over sequencing depth (Rapaport et al., 2013). However, large RNA-seq experiments remain the exception, rather than the rule at the moment.

The count-based nature of RNA-seq data prompted new development of statistical methods to call DEG. Despite the latter, differential gene expression analysis remains challenging due to lack of standard guidelines for experimental design, read processing, normalization and statistical analysis (Auer and Doerge, 2010; Auer et al., 2012). Currently, there is a bewildering number of methods for calling DEG. Several recent studies compared the relative performance of various DEG call methods using simulated and also real RNA-seq data sets (eleven in Soneson and Delorenzi (2013); five in Guo et al. (2013); eight in Seyednasrollah et al. (2015)), and offered recommendations for method selection. However, these studies did not explicitly consider variation of the performance of DEG call methods in the context biological effect size and unreplicated experiments, which are of practical concern to the biologist. It may not be an overstatement to say that, at present, how researchers pick a DEG call method out of the plethora of alternatives available is more guided by their degree of familiarity with the methodology literature, computing convenience and democratic evaluation of personal experiences in bioinformatics forums, rather than on empirical evidence.

Most DEG call methods are designed to address analysis of RNA-seq experiments that have biological replicates. A minority such as ASC (Wu et al., 2010), NOISeq (Tarazona et al., 2011) and GFOLD (Feng et al., 2012) were initially designed for analysis of unreplicated experiments, though the latter two could also handle replicated experiments. While unreplicated experiments are not suitable for reliable inference at the systems level, DEG mined using particular DEG call methods may nonetheless be useful for targeted study if their expression can be validated independently using qPCR. Such small incremental gains can be crucial to build up the ground work in preparation for more extensive study in non-model organisms. Our study aims to clarify the interaction between replicate size, biological effect size and DEG call method, so as to provide practical recommendations for RNA-seq data analysis that will help researchers get the most out of their RNA-seq experiments.

MATERIALS AND METHODS

Statistical methods for calling differentially expressed gene

A large number of DEG call methods have been proposed (Table 1), with the majority of them being parametric methods that make distributional assumption about the read count data. The relative performance of various subsets these methods have been investigated in several studies (Soneson and Delorenzi, 2013; Guo et al., 2013; Zhang et al., 2014; Seyednasrollah et al., 2015).

An exhaustive comparison of all available methods for the present study was not feasible nor necessary. Comparisons involving methods specifically designed for unreplicated experiments received little attention despite the abundance of such type of RNA-seq data. For this reason, we included GFOLD and ASC, both being Bayesian methods.

Method	Total citations	Citations per year	Reference
DESeq*	2987	597	Anders and Huber (2010)
edgeR*	2260	452	Robinson et al. (2010)
Cuffdiff2	517	258	Trapnell et al. (2013)
DESeq2*	209	209	Love et al. (2014)
voom*	143	143	Law et al. (2014)
DEGseq	592	118	Wang et al. (2010)
NOISeq*,+,n	324	81	Tarazona et al. (2011)
baySeq	310	62	Hardcastle and Kelly (2010)
SAMSeqn	114	57	Li and Tibshirani (2013)
EBSeq	107	53	Leng et al. (2013)
PoissonSeq	99	33	Li et al. (2012)
BitSeq	70	23	Glaus et al. (2012)
DSS	46	23	Wu et al. (2013)
TSPM	70	17	Auer and Doerge (2011)
NBPSeq	65	16	Di et al. (2011)
GFOLD*,+	44	15	Feng et al. (2012)
ShrinkSeq	30	15	Van De Wiel et al. (2013)
NPEBseqn	14	7	Bi and Davuluri (2013)
ASC*,+	32	6	Wu et al. (2010)
BADGE	2	1	Gu et al. (2014)

Table 1. Methods for calling differentially expressed genes in RNA-seq data analysis. Total citations were based on Google Scholar search result as of 22 September 2015, and normalized by number of years since formal publication. The methods were ranked according to their citations per year. Symbols: * for methods included in the present study; + for methods initially developed to analyze unreplicated RNA-seq data sets; n for non-parametric method. Programming language: C/C++ for GFOLD, Cuffdiff2 and BitSeq; Matlab for BADGE; R for the rest.

For replicated experiments, we focused on methods that have received the most attention from the scientific community (as reflected by their relatively high citations per year) - edgeR, DESeq and its new version DESeq2. These are parametric methods that explicitly model the distribution of count data using the negative binomial distribution. Initially designed for standard experiments with biological replicates, these methods were later modified to accommodate analysis of unreplicated experiments as well, but their performance relative to GFOLD and ASC remains unclear. We did not include two methods with high citations per year: Cuffdiff2 and DEGSeq, based on conclusions from recent method comparative analyses. For example, Cuffdiff2 was found to have very low precision when replicate size increased in the analysis of two large RNA-seq data sets from mouse and human (Seyednasrollah et al., 2015). Furthermore, Zhang et al. (2014) showed that edgeR had slightly superior performance in the receiver operating characteristic curve compared to DESeq and Cuffdiff2. Another comparative study involving DESeq, DEGseq, edgeR, NBPSeq, TSPM and baySeq showed that DEGseq had the largest false positive rate among them (Guo et al., 2013).

The inclusion of the popular non-parameteric method NOISeq provides a contrast between performance of parametric and non-parametric methods. Voom was included because it connects log-transformed read count data to the mature limma analysis pipeline (Smyth, 2004, 2005) that has been used so successfully for detecting DEG in microarray data analysis. Finally, the Z-test for equality of two proportions was included to set upper bounds in the tested performance metrics that are attainable by naive application of a common textbook statistical method. Let N_{ij} be the pooled normalized read counts of the ith gene in the jth phenotype class ($j = 1,2$), $N_{+j} = \sum_i N_{ij}$ the total normalized read counts in the jth phenotype class, and $N_{i+} = \sum_{j=1,2} N_{ij}$ the total normalized read counts of the ith gene in all phenotype classes. Specifically, the Z-test statistic for the ith gene is given by

$$Z_i = \frac{\hat{p}_{i1} - \hat{p}_{i2}}{\sqrt{\hat{p}_i(1-\hat{p}_i)/N}},$$

where $\hat{p}_{ij} = N_{ij}/N_{+j}$, $\hat{p}_i = N_{i+}/N$, and N is the total number of normalized counts. Table 2 provides a description of the core modeling approaches of the eight methods considered in the present study.

Method	Description	Reference
NOISeq	Non-parameteric modeling of odds of signal against noise; NOISeqBIO is a variant for handling replicated experiments which integrates the non-parametric framework of NOISeq with an empirical Bayes approach	Tarazona et al. (2011) Tarazona et al. (2015)
ASC	Empirical Bayes shrinkage estimation of log fold change	Wu et al. (2010)
GFOLD	Poisson count distribution; Bayesian posterior distribution for log fold change	Feng et al. (2012)
edgeR	Negative binomial count distribution; genewise dispersion parameter estimation via conditional maximum likelihood; empirical Bayes shrinkage of dispersion parameter; exact test for p-value computation	Robinson et al. (2010)
DESeq	Negative binomial count distribution; local regression modeling of mean and variance parameters	Anders and Huber (2010)
DESeq2	Negative binomial count distribution; generalized linear model; shrinkage estimation of dispersion parameter and fold change	Love et al. (2014)
voom	Estimates of mean-variance trend from log-transformed count data are used as input for the limma empirical Bayes analysis pipeline developed for microarray data analysis	Law et al. (2014)
Z-test	The Z-statistic for testing the equality of two proportions	-

Table 2. Description of the core modeling strategy of differential gene expression analysis methods investigated in the present study.

Criteria for differential expression

For edgeR, DESeq, DESeq2 and Z-test, we used a joint filtering criteria (Li, 2012) based on fold change (ϕ) and p-value (p) to call DEG. Let $y = -\log_{10} p$ and $x = \log_2 \phi$. Thus, each gene is associated with a paired score (x,y) after differential expression analysis. Following (Xiao et al., 2014), we required $p < 0.01$ and $\phi \geq 2$ to call for up-regulated genes, and $p < 0.01$ and $\phi \leq 1/2$ to call for down-regulated genes. The product of $y > 2$ and $|x| \geq 1$ yields the inequality $y > 2/|x|$. Thus, genes that fell in the region defined by $y > 2/x$ were differentially up-regulated, and those in the region of $y > -2/x$ were differentially down-regulated. The union of the sets of differentially up and down-regulated genes made up the set of DEG candidates.

For edgeR, we used the exact test option to perform differential expression analysis. To handle unreplicated experiments, we set the biological coefficient of variation (BCV) parameter as 0.4 for the Cheung data set (see details in Benchmarking section), and 0.1 for the Bottomly data set, following recommendations in Chen et al. (2015).

For NOISeq, we used the recommended criteria for calling DEG as described in the NOISeq documentation - $q = 0.9$ for unreplicated experiments, and $q = 0.95$ for experiments with biological replicates.

For ASC, we called DEG using double filtering of estimated $\log_2$ of fold change (FC) and estimated posterior probability, where $|\log_2 \mathrm{FC}| \geq 1$ and posterior probability $\geq 99\%$.

For GFOLD, we used the default significant cut-off of 0.01. A gene with GFOLD value of 1 or larger was considered differentially up-regulated, and differentially down-regulated if GFOLD value was -1 or smaller. Except GFOLD, which is written in the C/C++ language and requires the Linux platform, the other methods were executed in R version 3.1.3 (R Core Team, 2015).

Benchmarking

Data sets

To set up our benchmarking exercise, we needed two RNA-seq data sets whereby variation in their phenotype classes produced mild and strong biological effect sizes in the tissue of interest, respectively. We further required the RNA-seq data sets to have fairly large replicate sizes to enable the simulation of

different replicate size scenarios. To this end, we identified two suitable RNA-seq data sets in the Recount database (Frazee et al., 2011). The latter contains unnormalized RNA-seq count data sets from 18 major studies that have been assembled from raw reads using the Myrna (Langmead et al., 2010) pipeline.

The Bottomly data set (Bottomly et al., 2011) consists of gene expression data (22 million Illumina reads per sample, read length of ∼30 bases) obtained from the brain striatum tissues of two mice strains: C57BL/6J ($n = 10$) and DBA/2J ($n = 11$). Both mice strains are known to show large, strain-specific variation in neurological response when subjected to opiate drug treatment (Korostynski et al., 2006, 2007; Grice et al., 2007).

The Cheung data set (Cheung et al., 2010) consists of gene expression data (40 million Illumina reads per sample, read length of 50 bases) from immortalized human B-cells of 24 males and 17 females. Sex hormones are known to modulate B cell function (Klein, 2000; Verthelyi, 2001). For example, estrogen modulates B cell apoptosis and activation (Grimaldi et al., 2002), while testosterone suppresses immunoglobulin production by B cells (Kanda et al., 1996). In the absence of antigenic challenge, however, it seems reasonable to expect only a modest number of DEG in male and female B cells.

After removal of transcripts with zero counts in all samples, the Bottomly count table contained 13932 transcripts, down from an initial 36536 transcripts, whereas the Cheung count table contained 12410 transcripts, down from 52580. Prior to analysis, the count data were normalized using DESeq normalization (Anders and Huber, 2010), which has been shown to be robust to library size and composition variation (Dillies et al., 2013). However, raw counts were used for DESeq2 analysis since the method explicitly requires such type of data as input.

Method for constructing reliable reference DEG set

The construction of a reliable reference DEG set from which performance metrics for each method is evaluated is a non-trivial problem, if one eschews a simulation-based approach. To avoid circular reasoning, this reference set needs external validation from independent evidence such as confirmation from qPCR results.

Here, we chose voom (Law et al., 2014; Ritchie et al., 2015) as the method of choice for setting the reference DEG set. Unlike other DEG methods that primarily model mean-variance relationships in the count data using discrete distributions such as the Poisson or negative binomial distributions, voom log-transforms count data into a microarray-like data type suitable for analysis using the robust limma pipeline (Smyth, 2004; Ritchie et al., 2015). Because of this, using voom to set the reference DEG set can avoid biasing results of the called DEG due to algorithmic similarities. A gene was defined as differentially expressed using the same joint filtering criteria for edgeR, DESeq, DESeq2 and Z-test. We found the nonparametric SAMSeq (Li and Tibshirani, 2013), which has also been reported to have strong DEG mining performance, unsuitable for setting the reference DEG set as it returned different DEG sets for different random seeds and number of permutation parameters (Supplemental Material Fig. S1).

We assessed the validity of voom as a tool for constructing reasonable in silico reference DEG sets for the Bottomly and Cheung data set by comparing its performance with other DEG call methods on an RNA-seq data set in which sufficiently large numbers of qPCR-validated genes are available. Briefly, the Rajkumar data set (Rajkumar et al., 2015) consisted of gene expression count data (26119 genes; minimum of 10 million Illumina reads per sample, read length of $\sim$ 50 bases) from the amygdala tissue of C57BL/6NTac strain mice. There were two phenotype classes: wild type ($n = 8$), and heterozygotes for the *Brd1* gene deletion ($n = 8$). A total of 115 genes were selected for qPCR validation (Additional Table 5 in Rajkumar et al. (2015)); differential expression was observed in 60 of them, and not in the remaining 55. Each DEG call method returns N_g differentially expressed gene candidates. We considered a method to be sound for setting the reference DEG set if it did not return too few (tens) or too many (thousands) candidates. Among methods that satisfied this criterion, the one that had relatively higher positive predictive value (PPV; the complement of the false discovery rate) would be preferred. Let N_{TP} be the number of true positives, and N_{FP} the number of false positives. Then the number of DEG that lack validation result is $U = N_g - N_{FP} - N_{TP}$. If U is not too large or too small, then the expected number of true positives can be estimated as

$$N^*_{TP} = N_{TP} + \frac{N_{TP}}{N_{TP} + N_{FP}} U.$$

The expected PPV is therefore given by

$$PPV^* = \frac{N_{TP} + N^*_{TP}}{N_g}.$$

Characteristics of constructed reference DEG sets

Ideally, the in silico reference DEG set called using voom for the two test data sets should be independently validated using qPCR, but evidence at such level may not always be available. Where microarray data are available for the same study, a DEG candidate can be considered reliable if it is called in both RNA-seq and microarray analyses, since fold change of DEG from the latter has been found to correlate strongly with fold change from qPCR (Wang et al., 2014). A total of 362 DEG for the Bottomly data set were thus called (Fig. 1a). About 88% (320/362) of the DEG for the Bottomly data called using voom were identical with those called in Bottomly et al. (2011) using edgeR (1727 DEG). Approximately two fifths of them (153/362) were detected using limma applied on Affymetrix or Illumina microarray expression data (Supplemental Material Table T1).

For the Cheung data set, gender difference was the source of phenotype class variation. We exploited this fundamental biological difference to infer the most reliable DEG from the candidates returned using voom. Only DEG which were located on the sex chromosomes, or interacted with at least one gene product from the sex chromosomes were used to construct the reference DEG set. This strategy resulted in a set of 19 DEG (Fig. 1b). Five of them were located on the Y chromosome, three on the X chromosome and the remainder had known gene-gene interactions with at least one gene located on sex chromosomes (based on BioGRID (Stark et al., 2006; Chatr-Aryamontri et al., 2015); Supplemental Material Table T2).

Differentially expressed genes are characterized by between-phenotype variation that is significantly larger than within-phenotype variation. However, occasionally some genes may be wrongly declared as differentially expressed because some outliers within a phenotype class were sufficiently extreme to cause relatively large between-phenotype variation. To assess the quality of DEG called using voom, we used the Bland-Altman plot (Bland and Altman (1986); Supplemental Material Figure S2). Among the 362 DEG called for the Bottomly data set, the majority of DEG showed good agreement of replicate variation between the two phenotype classes - about 88% (318/362) were within 2SD (standard deviation) from perfect agreement, and about 95% (345/362) were within 3SD. Similarly, among the 19 DEG called for the Cheung data set, within-phenotype variation difference was within 2SD from perfect agreement for about 74% (14/19) of DEG, and within 3SD for about 84% (16/19) of DEG. Generally, genes that showed large within-phenotype variation in both phenotypes were not called by voom.

Once the DEG set had been constructed for the Bottomly and Cheung sets, it became possible to operationally define what we meant by mild or strong biological effect size. For the *i*th differentially expressed gene, define

$$T_i^2 = \frac{(\bar{X}_{i,1} - \bar{X}_{i,2})^2}{S^2_{i,1}/N_1 + S^2_{i,2}/N_2},$$

where i indexes the genes, and j the phenotype classes ($j = 1,2$); $\bar{X}_{i,j}$ and $S^2_{i,j}$ are the mean and variance of normalized read counts respectively, and n_j are the replicate sizes. Thus, T^2 is essentially the square of the t-statistic, which measures the magnitude of squared deviation between mean counts in two different phenotype classes relative to the latters' variances. By definition, the median values of T^2 should be large in a data set that shows strong biological effect size, and vice versa. For the Bottomly data set (strong effect size), median T^2 was 27.6; for the Cheung data set (mild effect size), it was 4.6. Both data sets had approximately equal spread of T^2 values around the median, the interquartile range being 38.3 and 34.5 for the Bottomly and Cheung data sets, respectively.

Simulation and performance evaluation

To simulate unreplicated experiments in both data sets, we used all possible sample pairs ($11 \times 10 = 110$) for the Bottomly data set, and 300 random sample pairs for the Cheung data set. Except ASC, which only handles unreplicated experiments, we further examined the behavior of other DEG call methods in cases of low to modest number of replicates. We constructed 100 instances of experiments for each replicate size per phenotype class in the Cheung data set ($n = 3,6,10$), and the Bottomly data set ($n = 3,6$) by random sampling without replacement within each phenotype class.

To evaluate method performance, we used sensitivity and positive predictive value (PPV). For each DEG call method, we computed sensitivity as the proportion of reference DEG that were called. PPV

was computed as the proportion of DEG called that were members of the reference DEG set. The mean and standard deviation (SD) of these metrics were then reported. Methods that show good PPV are particularly interesting in the context of unreplicated experiments, since DEG obtained from them offer the best potential of being validated. For systems level analysis, DEG should preferably be called using methods with good balance of sensitivity and PPV.

RESULTS & DISCUSSION

Validity of voom for setting the reference DEG set

The DEG set size and expected PPV of each method in the analysis of the Rajkumar data set are given in Table 3. The results indicate that only voom and edgeR produced call sizes that were of reasonable order of magnitude. However, voom had relatively higher expected PPV over edgeR; additionally, the DEG set size called using voom had standard error (SE) that was an order of magnitude smaller compared to edgeR (bootstrap sampling with replacement of biological replicates; 1000 iterations). Therefore, it seemed reasonable to use voom as the method of choice to construct the reference DEG set for the Bottomly and Cheung data sets.

Method	DEG set size $\pm$ SE	$PPV^{*}\pm$ SE (%)
voom	287 ± 43	88.9 ± 4.1
edgeR	564 ± 694	72.6 ± 15.0
DESeq	3384	NR
Z-test	9417	NR
DESeq2	10	NR
NOISeq	31	NR
GFOLD	38	NR

Table 3. DEG set size and expected PPV of the DEG call methods in the analysis of Rajkumar data set. Variation in DEG set size and expected PPV were computed using bootstrapping for methods where the DEG set size was not too small or too large. Abbreviation: SE for standard error; NR for "Not Relevant"

We note with interest from Table 3 that the number of DEG called by DESeq2 dropped drastically compared to DESeq. Since DESeq2 implements a shrinkage estimation of dispersion parameter and fold change to improve the performance of DESeq, the present suggests that this may occasionally lead to over-correction which results in DEG set size that is too small.

Performance of DEG call methods in the Cheung and Bottomly data sets

Positive predictive value and sensitivity

The ASC package provided by Wu et al. (2010) failed to run for particular combinations of sample pairs. Only 24.5% (27/110) and 41.3% (124/300) pairs of samples from the Bottomly and Cheung data set respectively could analyzed using ASC. The simulation results show that optimality of a DEG call method for a given replicate size depended on whether biological effect size was mild or strong (Fig.2). In the Cheung data set (mild biological effect size), all methods had very low (about 1%) mean positive predictive value (PPV) for unreplicated experiment (Fig. 2a; Supplemental Material Table T3), suggesting that no meaningful biological insights were possible. However, mean PPV ($\pm$SD) increased substantially for NOISeq to $43.5 \pm 31.5\%$, and for GFOLD to $29.6 \pm 15.8\%$, for $n = 3$ (Fig. 2b). Doubling and approximately tripling the latter to $n = 6$ (Fig. 2c) and $n = 10$ (Supplemental Material Fig. S3) further improved mean PPV for NOISeq to $87.0 \pm 16.1\%$ and $92.2 \pm 12.9\%$, and for GFOLD to $36.3 \pm 14.9\%$ and $52.6 \pm 18.8\%$, respectively. In all four replicate sizes, mean PPV was low for the other methods, not exceeding 12% for DESeq2, and was never more than 3% for edgeR, DESeq and Z-test.

A markedly different pattern of method performance was observed in the analysis of the Bottomly data set (strong biological effect size). In unreplicated experiments (Fig. 2d), mean PPV was relatively high for NOISeq ($50.6 \pm 20.3\%$), ASC ($47.2 \pm 25.9\%$) and GFOLD ($31.2 \pm 25.6\%$), compared to just about 15% in edgeR and 5% in DESeq and Z-test. DESeq2 did not perform well, with mean PPV ($29.6 \pm 28.4\%$), and an extremely low sensitivity ($0.2 \pm 0.6\%$) as a result of making too few calls. Interestingly, GFOLD attained very high mean PPV at $n = 3$ ($94.3 \pm 6.9\%$; Fig. 2e), with marginal change to $92.5 \pm 3.3\%$ at

$n = 6$. However, GFOLD was also the method with the lowest sensitivity (below 10%) under these two designs, which was caused by its small DEG set size (Fig.3).

DESeq2 struck the best balance between PPV and sensitivity as replicate size increased, but edgeR showed reasonable performance too. At $n = 3$ (Fig. 2e) and $n = 6$ (Fig. 2f), DESeq2 had mean PPV of $52.5 \pm 10.8\%$ and $62.1 \pm 7.7\%$, with mean sensitivity of $36.0 \pm 5.7\%$ and $65.1 \pm 4.5\%$, respectively. For edgeR, its mean PPV was 28.7 ± 4.1 and 33.9 ± 3.0, with mean sensitivity of 59.8 ± 5.4 and 79.0 ± 4.6, respectively. At $n = 6$, DESeq2 had similar sensitivity compared to its older version DESeq, and a superior mean PPV that was about four times higher. Unsurprisingly, the Z-test remained the worst performer, with mean PPV just about 6%.

The general increase in mean sensitivity for replicated experiments was consistent with the finding that the increase in statistical power for detecting DEG is primarily determined by biological replicate size, and less by sequencing depth (Liu et al., 2014).

DEG set size

Figure 3 shows the distribution of DEG set size in the Cheung and Bottomly data sets for different replicate sizes (for details, see Supplemental Table T3). Although DESeq2 could be used to call DEG for unreplicated experiments, it was shown to behave erratically in the Bottomly data set, with extremely low mean DEG set size (2.5 ± 6.8). In general, for replicated studies, methods such as DESeq2, DESeq, edgeR and Z-test made large numbers of calls that were typically one or two order of magnitudes more (depending on underlying biological effect size) compared to GFOLD or NOISeq. Consequently, it is expected that their sensitivity would increase at the expense of PPV.

Optimality requires a context

The current results suggest that unreplicated RNA-seq experiments, which are still very common among underfunded labs working with non-model organisms, may be a cost-effective way to generate candidate DEG with reasonable likelihood of being validated, provided that the underlying biological effect size is strong. Thus, for unreplicated RNA-seq experiments with phenotype classes such as those associated with pathogenic challenge and physico-chemical stress, we expect DEG called using NOISeq or GFOLD to be good candidates for validation. ASC may also be useful, though it should be noted that it could fail to run for particular combinations of sample pairs, as we found out in the present study. For validation work, GFOLD and NOISeq should be even more efficient once triplicates are available, but further replicate size increase produced only marginal mean PPV gain in the Bottomly data set, suggesting that using more than triplicates is not a cost-effective approach when validation of DEG candidates is the main research goal. When biological effect size is strong, we suggest that DESeq2 or edgeR as promising methods to mine DEG for systems biology work, on account of their good PPV and sensitivity balance. However, users should be aware of possibility that shrinkage estimation of dispersion and fold change procedure in DESeq2 may over-correct the initial estimates of these parameters, leading to a DEG set size that is too small, as discovered in the analysis of the Rajkumar data set (Table 3).

Research programs focusing on investigation of weak or modest biological effect sizes must have replicates, use NOISeq or GFOLD for DEG calling, and then to restrict the research goal to validation of the DEG candidates. Pursuing a systems biology (e.g. gene set analysis, functional enrichment) direction in such programs is not feasible, since in the Bottomly data set, the DEG set size of both GFOLD and NOISeq at $n = 10$ became too small (below 20).

Table 4 summarizes the recommended DEG call methods and research goals for the combinations of biological effect size and replicate size considered in the present study.

	Biological effect size	
Replicate size	Mild	Strong
1	nothing works	GFOLD^{v}, NOISeq^{v}
3+	GFOLD^{v}, NOISeq^{v}	GFOLD^{v}, DESeq2^{s}, edgeR^{s}

Table 4. Pragmatic DEG call methods for four combinations of biological effect size and replicate size, with suggested applications. Abbreviation: v for validation work; s for systems biology work.

Transcriptome coverage effect

Transcriptome coverage can be another important source of variation for the observed RNA-seq gene counts (Sims et al., 2014). Assuming transcriptome size was approximately equal for human and mouse, relative transcriptome coverage was about three times larger in the Cheung data set (human) compared to the Bottomly data set (mouse). Despite this, detection of DEG remained difficult when biological effect size was mild, suggesting that the effect of transcriptome coverage on DEG calling was probably marginal in the present study.

Limitations

In the present study, we justified the use of voom for setting the reference DEG set on the basis of its performance in the Rajkumar data set which has 115 qPCR-validated genes. Ideally, analyses of additional data sets of this nature would help us better understand the variability of method performance. Unfortunately, RNA-seq data sets that are coupled with extensive qPCR validation results remain uncommon.

The cost of not constructing the reference gene set using a simulation approach was the lost of one degree of freedom in the number of DEG call methods that could be evaluated, since we had to select one of the methods to determine the reference DEG set. Because of this, it may be possible that voom was actually the ideal method for making DEG calls when sufficient replicates are available. Therefore, in practical situations, one may wish to consider taking the union of DEG set called using voom with that from DESeq2 or edgeR. If the size of the union set is too large, one may consider taking the intersection set instead to obtain a smaller, but higher confidence DEG set (Zhang et al., 2014).

Future prospects

Many biologists have difficulty publishing results of RNA-seq experiments with no or few biological replicates. Despite including qPCR validation results, these studies are often dismissed by reviewers simply on grounds of 'not having enough sample size'. This stand is unnecessarily dogmatic, and does not take into account that some particular combinations in the trinity of replicate size-effect size-call method can potentially yield biologically meaningful results, as shown in the present study.

It is gradually being appreciated that RNA-seq analysis is a complex analysis that needs to address the numerous sources of variation from library preparation to bioinformatic processing (Kratz and Carninci, 2014) to yield an interpretable result. As a corollary, we suggest that one-size-fits-all pipelines for RNA-seq analysis commonly adopted by bioinformatics service providers should not be expected to always yield the most optimal set of DEG. There is a certainly a need for greater consultation between scientist and the bioinformatician to fine-tune pipelines by taking into account interactions in the replicate size-effect size-call method trinity.

As more high-quality RNA-seq experimental data continue to accrue in public databases, a better understanding of the anticipated behavior of various DEG calling methods under different biological and replicate size scenarios should gradually emerge from systematic comparison studies such as the current one. A complete dummy's guide to RNA-seq differential expression analysis may not be too far ahead in the future.

CONCLUSIONS

In RNA-seq experiments, biological effect size is an important determinant of whether a research program at the individual gene or systems level would yield the most biological insight. When it is expected to be mild, RNA-seq experiments should primarily aim at mining DEG for validation purpose, using at least triplicates and either NOISeq or GFOLD for DEG calling. Moreover, systems level analysis remains difficult as none of the methods considered presently showed satisfactory sensitivity and positive predictive value performance. When strong biological effect size is expected, analysis of unreplicated experiments using GFOLD or NOISeq can yield DEG candidates with optimistic validation prospects. A standard triplicate design (Liu et al., 2014) should result in further improvement. DESeq2 or edgeR seems to be most suited for calling DEG for subsequent systems level analysis as each showed good compromise between PPV and sensitivity. Combining results from voom with those from DESeq2 or edgeR may lead to further improvements.

ADDITIONAL INFORMATION AND DECLARATIONS

Competing Interests

The authors declare there are no competing interests.

Author Contributions

Tsung Fei Khang conceived, designed the experiments, analyzed the data and wrote the paper. Ching Yee Lau performed computational analyses, prepared figures and tables, analyzed the data and discussed the results.

Code Availability

R codes for the computational analyses done are available at http://github.com/tfkhang/rnaseq.

ACKNOWLEDGEMENTS

We are grateful to Jose M.G. Izarzugaza, Hao Zheng, an anonymous reviewer, and Jaume Bacardit (Academic Editor) for their helpful and constructive comments which resulted in important improvements to the present work.

FIGURES

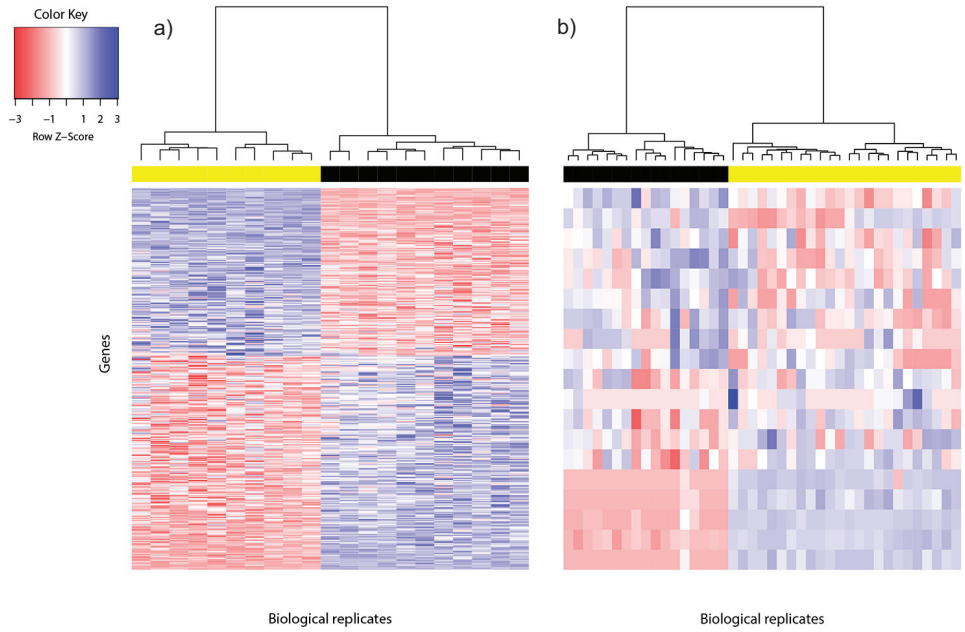


Figure 1. Heat map of differentially expressed genes in a) Bottomly data set (362 DEG) and b) Cheung data set (19 DEG). Phenotype class legend: a) Black for DBA/2J strain ($n = 11$); yellow for C57BL/6J strain ($n = 10$). b) Black for female ($n = 17$); yellow for male ($n = 24$). The heat maps were made using the `gplots` (Warnes et al., 2014) R package. Pairwise sample distances were estimated using the Euclidean distance and sample clustering was done using the Ward algorithm. The DEG were sorted based the magnitude and sign of their t-statistic.

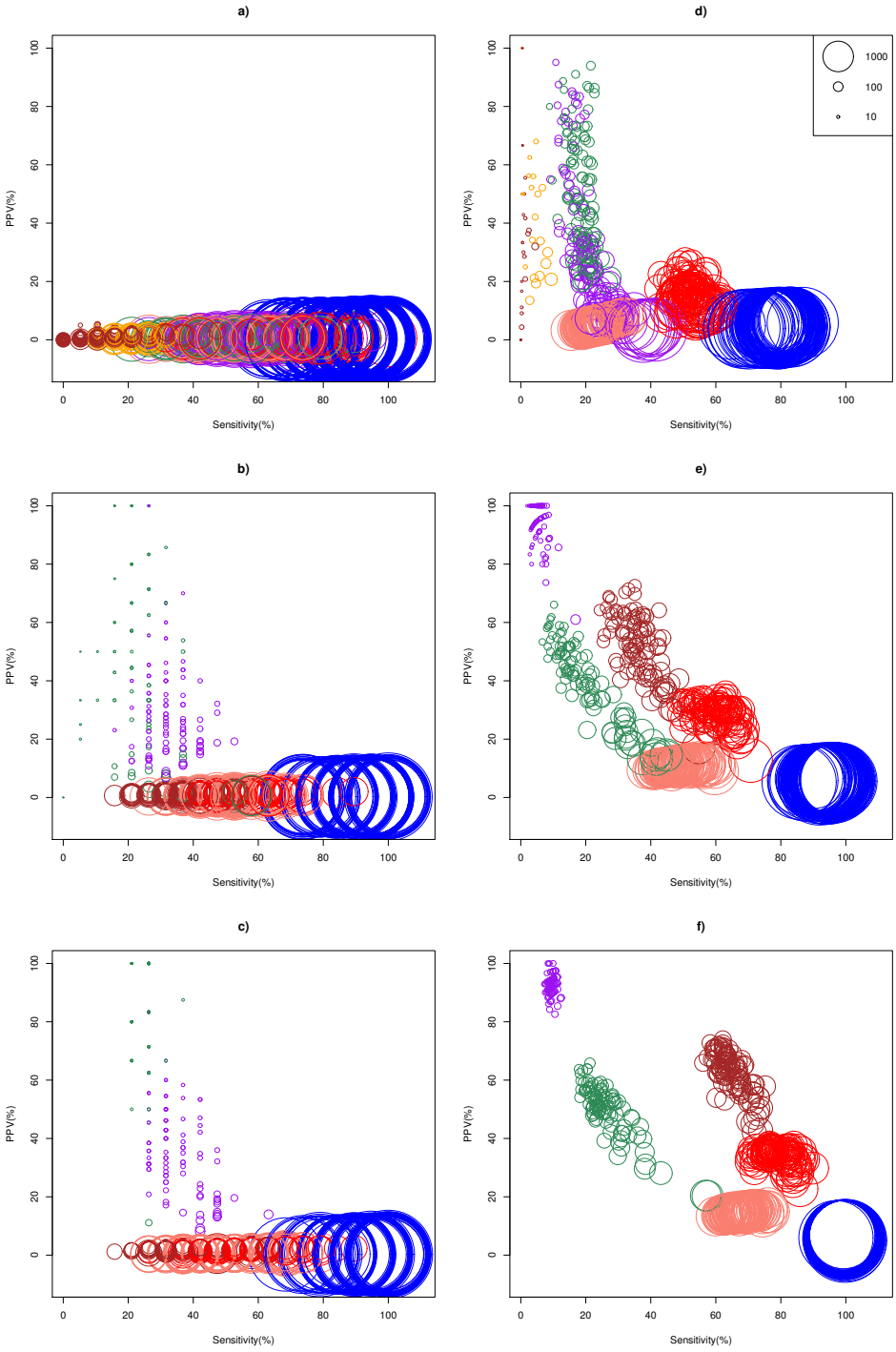


Figure 2. Scatter plots of PPV against sensitivity. The $n = 1, 3, 6$ scenarios are given in panels a,b,c for the Cheung data set, and d,e,f for the Bottomly data set, respectively. The diameter of a circle is proportional to the DEG set size (scale provided in Fig. 2d). Color legend: blue(Z-test), pink(DESeq), red(edgeR), brown(DESeq2), purple(GFOLD), green(NOISeq), orange(ASC). For $n = 10$ in the Cheung data set, see Supplemental Material Fig. S3.

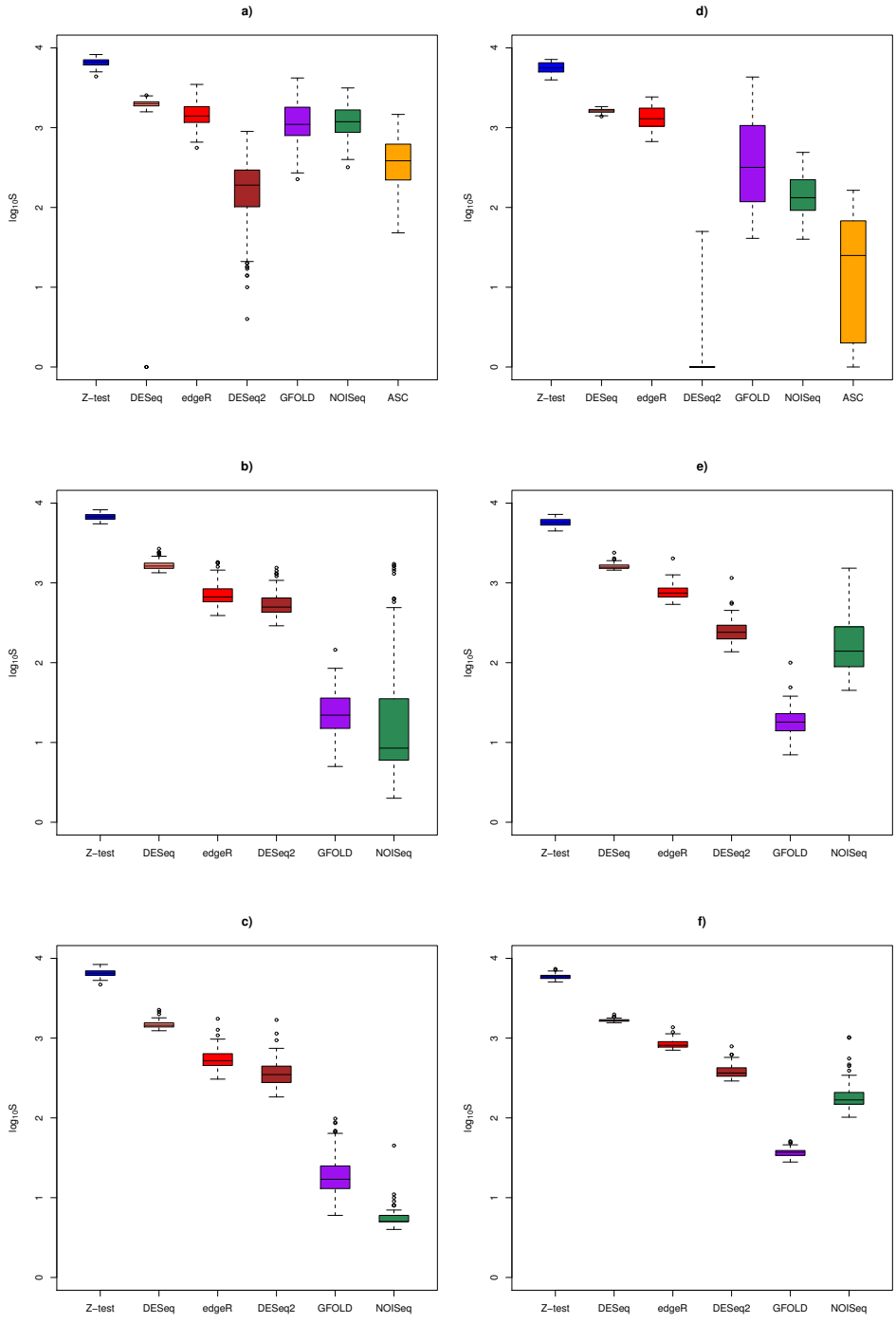


Figure 3. Box plots of distribution of DEG set size (in $\log_{10}$ scale) by method. The $n = 1, 3, 6$ scenarios are given in panels a,b,c for the Cheung data set, and d,e,f for the Bottomly data set, respectively. Color legend: blue(Z-test), pink(DESeq), red(edgeR), brown(DESeq2), purple(GFOLD), green(NOISeq), orange(ASC). For $n = 10$ in the Cheung data set, see Supplemental Material Fig. S4.

REFERENCES

Allison, D. B., Cui, X., Page, G. P., and Sabirpour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7:55–65.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11:R106.

Auer, P. L. and Doerge, R. W. (2010). Statistical design and analysis of RNA sequencing data. *Genetics*, 185:405–416.

Auer, P. L. and Doerge, R. W. (2011). A two-stage Poisson model for testing RNA-seq data. *Statistical Applications in Genetics and Molecular Biology*, 10:1–26.

Auer, P. L., Srivastava, S., and Doerge, R. (2012). Differential expression - the next generation and beyond. *Briefings in Functional Genomics*, 11:57–62.

Bi, Y. and Davuluri, R. v. (2013). NPEBseq: nonparameteric empirical Bayesian-based procedure for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14:262.

Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327:307–310.

Bottomly, D., Walter, N. A., Hunter, J. E. Darakijan, P., Kawane, S., Buck, K. J., Searles, R. P., Mooney, M., McWeeney, S. K., and Hitzemann, R. (2011). Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLoS One*, 24:e17820.

Chatr-Aryamontri, A., Breitkreutz, B. J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., Reguly, T., Nixon, J., Ramage, L., Winter, A., Sellam, A., Chang, C., Hirschman, J., Theesfeld, C., Rust, J., Livstone, M., Dolinski, K., and Tyers, M. (2015). The BioGRID interaction database: 2015 update. *Nucleic Acids Research*, 43:D470–D478.

Chen, Y., McCarthy, D., Robinson, M., and Smyth, G. (2015). edgeR: differential expression analysis of digital gene expression data. `http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf`. [Online; accessed 27-May-2015].

Cheung, V. G., Nayak, R. R., Wang, I. X., Elwyn, S., Cousins, S. M., Morley, M., and Spielman, R. S. (2010). Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biology*, 8:e1000480.

Di, Y., Schafer, D. W., Cumbie, J. S., and Chang, J. H. (2011). The NBP negative binomial model for assessing differential gene expression from RNA-seq. *Statistical Applications in Genetics and Molecular Biology*, 10:1–28.

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmouqin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jaqla, B., Journeau, L., Laloö, D., Le Gall, C., Schaëffer, B., Le Crom, S., Guedj, M., Jaffrézic, F., and Consortium, F. S. (2013). A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14:671–683.

Dudoit, S., Shaffer, J. P., and Boldrick, L. (2014). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18:71–103.

Feng, J., Meyer, C. A., Wang, Q., Liu, J. S., Liu, X. S., and Zhang, Y. (2012). GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics*, 28:2782–2788.

Frazee, A. C., Langmead, B., and Leek, J. T. (2011). Recount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, 12:449.

Glaus, P., Honkela, A., and Rattray, M. (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, 28:1721–1728.

Grice, D. E., Reenilä, I., Männistö, P. T., Brooks, A. I., Smith, G. G., Golden, G. T., Buxbaum, J. D., and Berrettini, W. H. (2007). Transcriptional profiling of C57 and DBA strains of mice in the absence and presence of morphine. *BMC Genomics*, 8:76.

Grimaldi, C. M., Cleary, J., Selma Dagtas, A., Moussai, D., and Diamond, B. (2002). Estrogen alters thresholds for B cell apoptosis and activation. *The Journal of Clinical Investigation*, 109:1625–1633.

Gu, J., Wang, X., Halakivi-Clarke, L., Clarke, R., and Xuan, J. (2014). BADGE: a novel Bayesian model for accurate abundance quantification and differential analysis of RNA-Seq data. *BMC Bioinformatics*, 15(Suppl 9):S6.

Guo, Y., Li, C. I., Ye, F., and Shyr, Y. (2013). Evaluation of read count based RNAseq analysis methods. *BMC Genomics*, 14(Suppl 8):S2.

Hardcastle, T. J. and Kelly, K. A. (2010). baySeq: empirical Bayesian methods for identifying differential

expression in sequence count data. *BMC Bioinformatics*, 11:422.

Jeanmougin, M., de Reynies, A., Marisa, L., Paccard, C., Nuel, G., and Guedj, M. (2014). Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. *PLoS One*, 5:e12336.

Kanda, N., Tsuchida, T., and Tamaki, K. (1996). Testosterone inhibits immunoglobulin production by human peripheral blood mononuclear cells. *Clinical & Experimental Immunology*, 106:410–415.

Klein, S. L. (2000). The effects of hormones on sex differences in infection:from genes to behavior. *Neuroscience & Biobehavioral Reviews*, 24:627–638.

Korostynski, M., Kaminska-Chowaniec, D., Piechota, M., and Przewlocki, R. (2006). Gene expression profiling in the striatum of inbred mouse strains with distinct opiod-related phenotypes. *BMC Genomics*, 7:146.

Korostynski, M., Piechota, M., Kaminska, D., Solecki, W., and Przewlocki, R. (2007). Morphine effects on striatal transcriptome in mice. *Genome Biology*, 8:R128.

Kratz, A. and Carninci, P. (2014). The devil in the details of RNA-seq. *Nature Biotechnology*, 32:882–884.

Langmead, B., Hansen, K. D., and Leek, J. T. (2010). Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biology*, 11:R83.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15:R29.

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M., and Kendziorski, C. (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, 29:1035–1043.

Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J. A., Stewart, R., and Dewey, C. N. (2014). Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biology*, 15:553.

Li, J. and Tibshirani, R. (2013). Finding consistent patterns:a nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research*, 22:519–536.

Li, J., Witten, D. M., Johnstone, I. M., and Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, 13:523–528.

Li, W. (2012). Volcano plots in analyzing differential expressions with mRNA microarrays. *Journal of Bioinformatics and Computational Biology*, 10:1231003.

Liu, Y., Zhou, J., and White, K. P. (2014). RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, 30:301–304.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550.

Oshlack, A., Robinson, M. D., and Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biology*, 11:220.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing., Vienna, Austria.

Rajkumar, A. P., Qvist, P., Lazarus, R., Lescai, F., Ju, J., Nyegaard, M., Mors, O., Børglum, A. D., Li, Q., and Christensen, J. (2015). Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. *BMC Genomics*, 16:548.

Rapaport, J., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D., and Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, 14:R95.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43:e47.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140.

Seyednasrollah, F., Laiho, A., and Elo, L. L. (2015). Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*, 16:59–70.

Sims, D., Sudberry, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15:121–132.

Smyth, G. (2005). limma: Linear models for microarray data. In Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., and Huber, W., editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 1.

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14:91.

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34:D535–539.

Tarazona, S., Furío-Tarí, P., Turrá, D., Di Pietro, A., Nueda, M. J., Ferrer, A., and Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with noiseq r/bioc package. *Nucleic Acids Research*, page doi:10.1093/nar/gkv711.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome Research*, 21:2213–2223.

Tibshirani, R. (2006). A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics*, 7:106.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., and Rinn, J. L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, 31:46–53.

Van De Wiel, M. A., Leday, G. G., Pardo, L., Rue, H., Van Der Vaart, A. W., and Van Wieringen, W. N. (2013). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, 14:113–128.

Verthelyi, D. (2001). Sex hormones as immunomodulators in health and disease. *International Immunopharmacology*, 1:983–993.

Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., Meehan, J., Li, X., Yang, L., Li, H., Labaj, P. P., Kreil, D. P., Megherbi, D., Gaj, S., Caiment, F., van Delft, J., Kleinjans, J., Scherer, A., Devanarayan, V., Wang, J., Yang, Y., Qian, H.-R., Lancashire, L. J., Bessarabova, M., Nikolsky, Y., Furlanello, C., Chierici, M., Albanese, D., Jurman, G., Riccadonna, S., Filosi, M., Visintainer, R., Zhang, K. K., Li, J., Hsieh, J.-H., Svoboda, D. L., Fuscoe, J. C., Deng, Y., Shi, L., Paules, R. S., Auerbach, S. S., and Tong, W. (2014). The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nature Biotechnology*, 32:926–932.

Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26:136–138.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10:57–63.

Warnes, G., Bolker, B. Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., and Venables, B. (2014). *gplots:Various R programming tools for plotting data. R package version 2.13.0.* http://CRAN.R-project.org/package=gplot.

Wu, H., Wang, C., and Wu, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, 14:232–243.

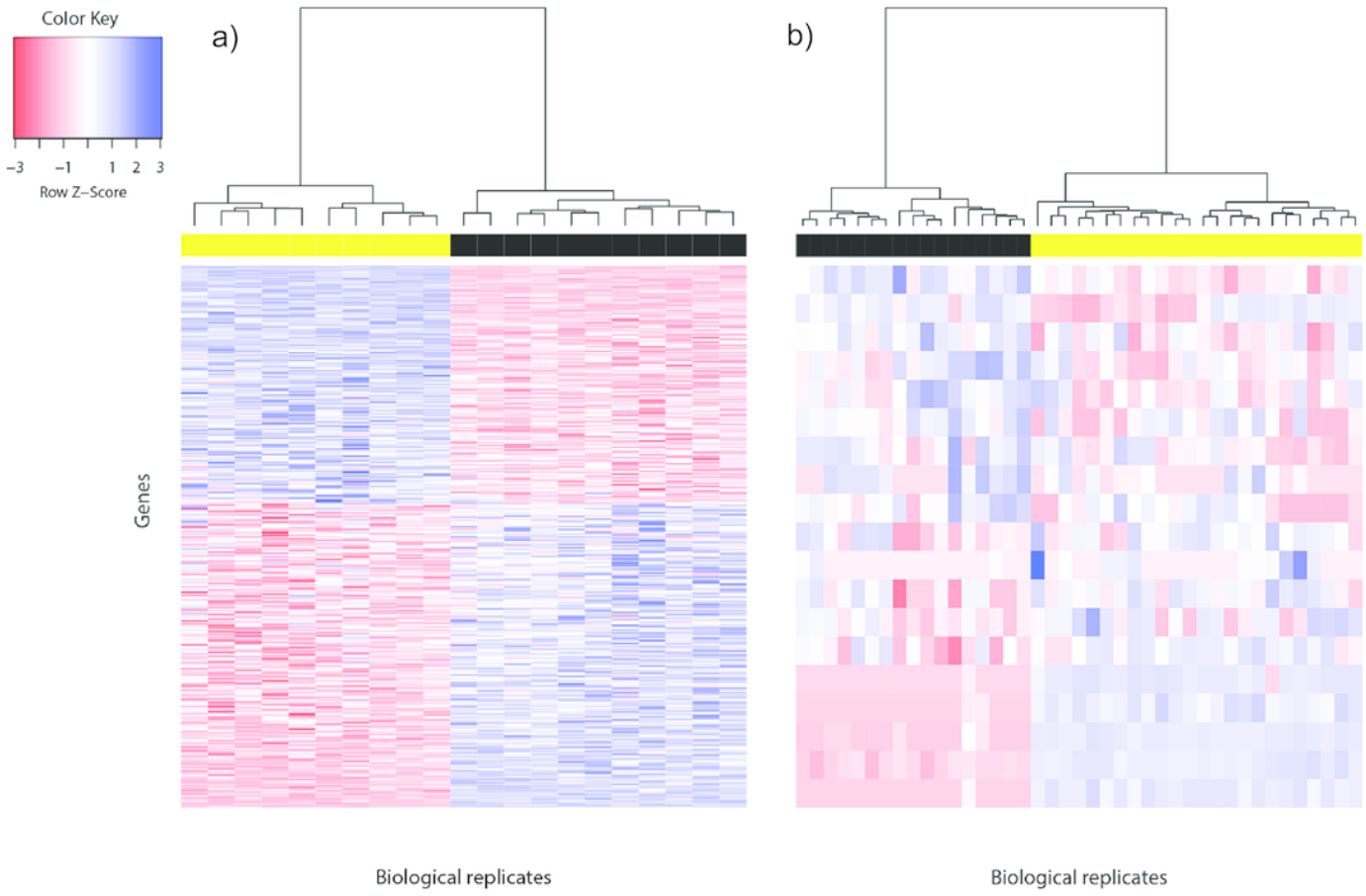
Wu, Z., Jenkins, B. D., Rynearson, T. A., Dyhrman, S. T., Saito, M. A., Mercier, M., and Whitney, L. P. (2010). Empirical Bayes analysis of sequencing-based transcriptional profiling without replicates. *BMC Bioinformatics*, 11:564.

Xiao, Y., Hsiao, T. H., Suresh, U., Chen, H. I. H., Wu, X., Wolf, S. E., and Chen, Y. (2014). A novel significance score for gene selection and ranking. *Bioinformatics*, 30:801–807.

Zhang, Z. H., Jhaveri, D. J., Marshall, V. M., Bauer, D. C., Edson, J., Narayanan, R. K., Robinson, G. J., Lundberg, A. E., Bartlett, P. F., Wray, N. R., and Zhao, Q. Y. (2014). A comparative study of techniques for differential expression analysis of RNA-seq data. *PLoS ONE*, 9:e103207.

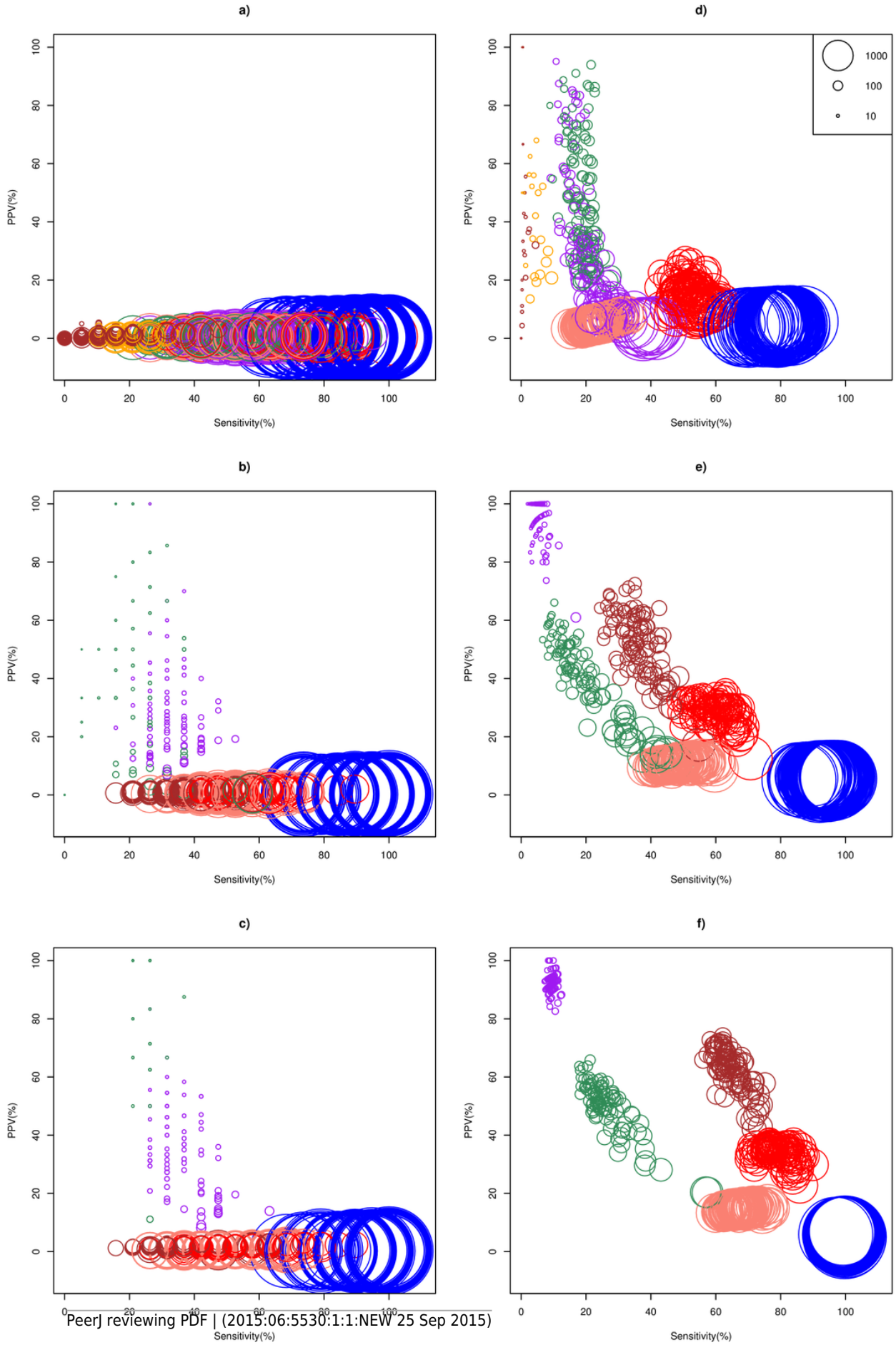
1

Figure 1


Color Key
−3 −1 1 2 3
Row Z−Score
a)
b)
Genes
Biological replicates
Biological replicates

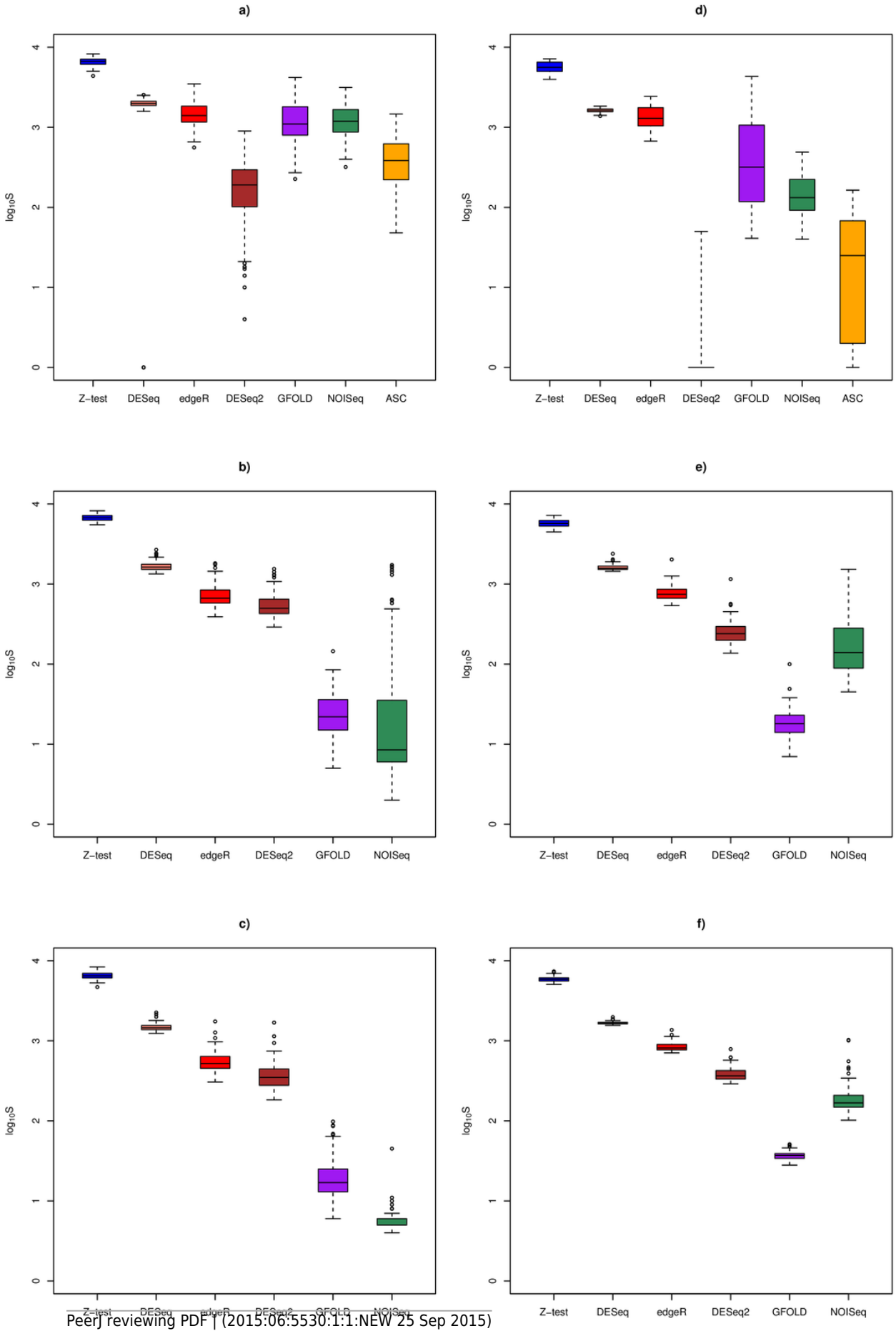
2

Figure 2


a)
d)
b)
e)
c)
f)
PPV(%)
Sensitivity(%)
0
20
40
60
80
100
1000
100
10

3

Figure 3


a)
d)
b)
e)
c)
f)
log10S
Z−test
DESeq
edgeR
DESeq2
GFOLD
NOISeq
ASC