Side-by-side analysis of alternative approaches on multi-level RNA-seq data

IRINA MOHORIANU[1,2],

[1]School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, United Kingdom.

[2]School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, United Kingdom.

[#]Corresponding author: i.mohorianu@gmail.com

Running title: Alternative approaches for multi-level RNA-seq data

ABSTRACT

Background:

RNA sequencing (RNA-seq) is widely used for RNA quantification across environmental, biological and medical sciences; it enables the description of genome-wide patterns of expression and the deduction of regulatory interactions and networks. The aim of computational analyses is to achieve an accurate output, i.e. rigorous quantification of genes/transcripts to allow a reliable prediction of differential expression (DE), despite the variable levels of noise and biases present in sequencing data. The evaluation of sequencing quality and normalization are essential components of this process.

Results:

We investigate the discriminative power of existing approaches for the quality checking of mRNA-seq data and also propose additional, quantitative, quality checks. To accommodate the analysis of a nested, multi-level design using data on *D. melanogaster*, we incorporated the sample layout into the analysis. We describe a "subsampling without replacement"-based normalization and identification of DE that accounts for the experimental design i.e. the hierarchy and amplitude of effect sizes within samples. We also evaluate the differential expression call in comparison to existing approaches**.** To assess the broader applicability of these methods, we applied this series of steps to a published set of *H. sapiens* mRNA-seq samples.

Conclusions:

The dataset-tailored methods improved sample comparability and delivered a robust prediction of subtle gene expression changes. Overall, the proposed approach offers the potential to improve key steps in the analysis of RNA-seq data by incorporating the structure and characteristics of biological experiments into the data analysis.

Keywords: RNA-seq; quality check; normalization; subsampling normalization; identification of differential expression; hierarchical differential expression.

Background

RNA sequencing (RNA-seq) has revolutionized the field of trascriptomics (Wang et al. 2009; Ozsolak et al. 2011), giving powerful insight into the identity and abundance of RNAs in cells, tissues and whole organisms (Jia et al. 2017). In contrast to the fixed, predefined set of probes used for microarray experiments, RNA-seq generates a diverse set of reads and facilitates analyses of expression level variation for known and unknown RNA transcripts and variants. It also offers the possibility to study additional facets of the transcriptome (Conesa et al. 2016a), such as the (re)annotations of the reference genomes (Torres-Oliva et al. 2016), the identification of alternative splicing events (Trapnell et al. 2012) and variation in abundance across transcripts (Dillies et al. 2013b; Patro et al. 2017).

Several bioinformatics methods have been used to analyse the rapidly rising number of RNA-seq datasets (reviewed in (Dillies et al. 2013b; Conesa et al. 2016a; Evans et al. 2017)). However, to accommodate the use of RNA-seq in complex experimental designs, there is scope for further developments in: (i) the robust detection of subtle signatures of gene expression (the concordance of which is often very low between different bioinformatics methods, (Rapaport et al. 2013; Roca et al. 2017), (ii) the incorporation of hierarchical experimental designs (Love et al. 2014b; Robinson et al. 2015; Schurch et al. 2016), e.g. from evolutionary experiments (Fang et al. 2011), and (iii) minimising the effect of normalisation on the pattern of DE, especially when differences are subtle (Dillies et al. 2013b). We discuss these themes in the major steps of RNA-seq analysis, below.

Quality Checks (QC)

A key step in the analysis of RNA-seq data consists of sample quality checks and the identification, characterization and potential exclusion of sample outliers, e.g. those samples that are compromised due to technical issues (Consortium 2014). Existing tools, such as FastQC (Andrews 2010), SeqMonk (Andrews 2010) or TagCleaner (Schmieder et al. 2010) focus on evaluating the sequencing and the per-base quality. Additional QC may include an analysis of the per-base nucleotide composition and an evaluation of the overall GC content

(Risso et al. 2011; DeLuca et al. 2012; Wang et al. 2012). QC procedures focused on the sequencing bias comprise the characterization of k-mer distributions (Hansen et al. 2010) as well as other adapter ligation effects, such as the secondary RNA structure of the insert with the attached adapters (Jayaprakash et al. 2011b; Jackson et al. 2014).

Quantitative analysis of sequencing output currently considers measures such as yield, coverage, 3'/5' bias, number of detectable transcripts, strand specificity and read distribution across the genome (Consortium 2014). Additional steps, at the transcript level, include the classification of reads into annotation classes, which can highlight the presence of potential contaminant ncRNAs such as tRNAs and rRNAs (DeLuca et al. 2012; Wang et al. 2012). Such analyses can reveal over-represented classes of sequences, which could be removed in order not to distort the subsequent normalization and lead to changes in the ranking of abundances.

In recognition of the multiple sources of variation present in a biological experiment, a QC criterion is based on the Pearson Correlation Coefficient (PCC) between the gene expressions in biological replicates for transcripts that are detected in both samples (McIntyre et al. 2011; Gierlinski et al. 2015). Values of $r^2 = 0.92$-0.98 are generally accepted. If the PCC falls below 0.9 the suggestion is to identify and potentially exclude the problematic samples (Conesa et al. 2016b). However, this criterion may lack discriminatory power; due to the high number of data points (expressed genes, e.g. vector containing >15K genes for *D. melanogaster*) correlations will often be very high between *all* samples, regardless of their quality.

Additional steps for the quantitative evaluation of samples are possible, but are, as yet, under-utilised. These techniques include analyses of per-sample or per-gene complexities defined as the ratio of non–redundant (NR, unique) to redundant (R, total) reads (Mohorianu et al. 2011a), and similarity comparisons (Jaccard 1901), (Beckers et al. 2017).

Normalization

The next key stage in the analysis of RNA-seq data is the normalization of gene expression levels (Mortazavi et al. 2008b; Conesa et al. 2016a; Roca et al. 2017). Initial reports describing RNA-seq suggested that no normalization method was required (Wang et al. 2009). However, subsequent studies correctly highlighted normalization as a critical step (Bullard et al. 2010; Dillies et al. 2013a). Normalization is designed to transform the distributions of abundances for each sample, without distortion, into distributions that can be compared. A good normalization increases the chances of an accurate call of DE. It accounts for differences in sequencing depths and in biases arising from the library preparation or its sequencing (Li et al. 2014; Li et al. 2015; Lin et al. 2016).

However, despite extensive attention from the community (Aanes et al. 2014), there is as yet no clear consensus on whether there is any single optimal normalization method (Dillies et al. 2013b; Roca et al. 2017). Nor is there any general appreciation of the potential magnitude of the consequences of ineffectively normalizing data; their extent depends on the amplitude and distribution of DE, with small gene expression differences being more sensitive than larger ones to the method of normalization. Therefore, particularly for analyses of subtle gene expression differences, it can be important to assess how well the data are normalized by a specific method (Wagner et al. 2012). Such tests are not a routine part of bioinformatics analyses (Evans et al. 2017).

Identification of Differential Expression

The goal of transcriptomics analyses is the accurate and unbiased identification of expressed genes and genes showing DE between treatments. The majority of existing methods exhibit a good level of overlap in terms of highly differentially expressed genes (Soneson et al. 2013; Roca et al. 2017) . However, their agreement is far less when DE is subtle. Comparative analyses of existing normalization procedures on real and simulated data sets show that only ~50% of significantly differentially expressed genes are identified by all methods (Rapaport et al. 2013; Lin et al. 2016) .

ANOVA-based methods are a powerful and extensively applied approach for the analysis of microarray data (Cui et al. 2003). However, such methods are based on *a priori*, to some extent arbitrary, significance thresholds and the type of experiment can greatly influence the expected number of genes showing DE. For example, if the DE frequency distribution is narrow, stringent p-value thresholds can indicate as statistically significant genes that show very small fold change differences. Therefore, the set of DE genes identified by a fixed p-value threshold may not necessarily reflect biologically important facets of the data. Such differences are unlikely to be validated by low throughput methods (Evans et al. 2017).

Newer methods such as DESeq2 (Love et al. 2014a) and edgeR (Robinson et al. 2010) are based on the negative binomial distribution model for expression levels, using the variance and mean linked by local regression to detect DE genes (DESeq2) and empirical Bayes methods for moderating the degree of over dispersion across transcripts (edgeR). However, existing methods do not easily accommodate inherent, hierarchical experimental design with variable amplitude of DE between the hierarchy levels.

To fully encompass this type of experiment (exemplified using the *D. melanogaster* dataset) we incorporated the structure and magnitude of gene expression differences into the analysis, prior to the DE call. We accounted for the specific type of variation in expression for behavioural experiments and structured the analysis framework for the mRNA-seq data (Fig. S1); from QC (including existing and new approaches) to a subsampling without replacement-based normalization and finally a hierarchical approach for the identification of transcripts showing DE. The proposed analysis also features the use of an adjustable, empirically-determined offset to filter out low abundance genes and a DE call using maximal confidence intervals.

The adapted methods also performed well in direct comparisons with existing approaches for the analysis of a publicly available *H. sapiens* mRNA-seq dataset. Overall, the methods may offer advantages in the analysis of complex, challenging datasets and are complementary to existing approaches.

RESULTS and DISCUSSION

To develop and test the adapted methods we used a *D. melanogaster* dataset in which we tested for subtle effects on gene expression in males exposed to mating rivals (Mohorianu et al. 2017). We compared the output of our pipeline with that of existing methods on the same input data. The description of the steps of this analysis, in line with the approaches described in (Conesa et al. 2016a), are presented in Fig S1. We also assessed the applicability of the adapted normalization on a publicly available *H. sapiens* mRNA-seq dataset.

Quality Checking

Stage 1: QC of sequencing quality

The *D. melanogaster* data comprised of 3 replicate mRNA-seq samples of 2 rival treatments (rivals versus no rivals), 2 body parts (Head+Thorax (HT) and Abdomen (A)) and 3 time exposure treatments (2h, 26h or 50h). The first stage of the QC on these data (Fig. S1) focussed on existing approaches (i) the analysis of FastQ quality scores (Andrews 2010), (ii) sequencing depth, (iii) nucleotide (nt) composition / GC content (DeLuca et al. 2012; Wang et al. 2012), (iv) strand bias and (v) proportions of genome and annotation classes - matching reads e.g. mRNAs, t/rRNAs, miRNAs, UTRs, introns, intergenic regions (Conesa et al. 2016a). The FastQ QC indicated good quality reads for all 50nt, though we observed high variability in sequencing depths. Variation in nucleotide content was observed across the first 12nt (Jayaprakash et al. 2011a), but was consistent after that with nucleotide composition of the *D. melanogaster* transcriptome. Strand bias was comparable across samples and the proportion of genome-mapping reads was high. Based on these stage 1 quality checks, all samples were retained for further analyses and entered stage 2 QC. The detailed results, supporting the conclusion that the samples were consistent based on these criteria, are presented in Tables S1 and S2 and in (Mohorianu et al. 2017).

Stage 2: Quantitative QC of replicate and sample comparability

Jaccard/intersection analysis

We computed the Jaccard similarity (Mohorianu et al. 2011b; Beckers et al. 2017) at the gene level, to compare the similarity in expression of the top 1000 most abundant genes present in each sample (Table S3). This measure evaluated what proportion of the most abundantly expressed genes in one sample are also the most abundant in the next sample, and so on. Being calculated on the top most abundant genes it is not biased by different numbers of expressed genes in each sample. In addition, by selecting ~5% of the most abundant genes, this measure is not biased by noise-derived variability in expression. Samples drawn from the same body parts shared > 90% similarity, and between body parts (HT versus A) the similarity dropped to ~50%. Similarity between the experimental (± rivals) treatments was sometimes higher than between replicates, which highlighted the need for alternative approaches for normalizing the gene expression levels.

Complexity analysis

A comparison of complexity (calculated as the ratio of non-redundant (NR), unique, reads to redundant (R), all, reads, on all mRNA matching reads) is an informative measure of the number of unique reads present in each sample, average abundance of reads and subsequently on the average coverage. It is also useful for identifying sample/replicate outliers. A complexity of ~0 would indicate a sample in which all reads were the same and 1 a sample where every single read was different. Sample complexity is influenced by sequencing depth. Samples with high sequencing depth have a lower overall complexity and vice versa. However, samples with comparable sequencing depths, but very different number of unique reads, may highlight incomparable replicates/ samples. To understand how this may influence the accuracy of DE, we calculated the variation in complexity, between replicates, at gene level (Fig. 1). This revealed sizeable differences in complexity for most genes (Fig. 1(A)), which correlate well with the presence of highly variable number of spurious incident reads, especially for low abundance genes; this conclusion was also confirmed by the point-to-point correlation, described in the next section. To reduce this technical variation and normalize the data we tested the effect of subsampling with- (Fig. 1(B)) and without- (Fig. 1(C)) replacement for each set of reads to a fixed total (see below). We observed a reduction in complexity as a result of the subsampling both with and without replacement and an increased similarity between replicates/ samples. However, the subsampling with replacement artificially indicated that the third replicate (R3) is acceptable (Fig. 1(B)), while the subsampling without replacement (Fig 1(C)) maintains the conclusion that the third replicate is an outlier.

Correlation analyses

Correlations were first calculated between gene expression vectors in each sample to assess their comparability (using Pearson (PCC), Spearman (SCC) and Kendall (KCC) correlation coefficients), similarly as in (Gierlinski et al. 2015). Using the raw expression levels, correlations were computed between each sample and every other. Correlations between HT and A samples (in the range of 0.75-0.8) were lower than correlations between same body part samples. This was expected on the basis of HT- and A-specific genes whose expression is restricted to each body part. When only A or HT samples were considered, all correlation coefficients were above 0.95 (Fig. S2). These results showed a high correlation between samples and no sample would be excluded as clear outlier, even though, based on other quantitative QC measures, it would be prudent to do so (see below). Hence, standard correlation metrics may not be sufficiently sensitive to evaluate sample quality.

Point-to-Point Correlation analyses

To gain additional insight we introduced the ‘point-to-point’ Pearson correlation coefficient (p2pPCC) for each gene, which is a standard Pearson Correlation computed across the whole length of a gene on its expression profile. Although we expect to see alternative splicing events between the different replicates and samples, the number of genes affected by this is expected to be small. Next, we evaluated the distribution of PCC for pairwise sample or replicate comparisons against the corresponding gene abundances. This analysis revealed a higher variability at low in comparison to high abundances and a knock-on effect on the DE among replicates, i.e. more DE at low in comparison to high abundances (Fig. S3).

The p2pPCC was also used to determine a variability threshold (noise threshold, denoted as “offset” in the DE call section). This value was calculated, for the whole dataset, once the abundances of the sequencing reads were normalized. Briefly, low abundance genes have a limited number of incident fragments that align to random locations on the gene. As the gene abundance increases, the alignment pattern of the incident fragments starts to resemble the gene model leading to an increase in the p2pPCC. We determined the offset as the gene abundance for which the median of the p2pPCC distribution was > 0.7 between replicates. A similar approach, based on the entropy of strand bias, was implemented for sRNA sequencing in (Beckers et al. 2017).

Overall the qualitative QC metrics were informative but lacked discriminatory power. The quantitative QC metrics focused on the comparability at gene level by analyses of complexity and similarity and represent a valuable addition to overall QC.

Subsampling based Normalizations

To attenuate the effect of variation in sequencing depth between replicates and samples described above, we implemented a subsampling (without replacement) normalization on read expression levels adapted from (Li et al. 2013b), enriched with several additional checks on the consistency of the subsample and the similarity to the original sample. The subsampling approaches already in use either focus on the subsampling with replacement option applied on the aligned reads (Gierlinski et al. 2015) or on gene abundances (Robinson et al. 2014), or the subsampling without replacement applied on all reads, without additional per-gene consistency checks (Stupnikov et al. 2015).

Incremental subsampling (without replacement) and bootstrapping-based sample checking

First we tested the homogeneity of each sample, as indicated in (Stupnikov et al. 2015). To evaluate the existence of high abundance reads which, due to their higher probability of being selected, could distort the normalized distributions, we conducted a subsampling exercise, from 95% down to 45% (in steps of 5%) of the original redundant reads. Part of the novelty of our approach is to assess the consistency of the sample by checking if the proportion of redundant genome matching reads was affected by the subsampling (Table S4 and Fig S4). Even when the data were subsampled to 45% of the original sequencing depth, the proportion of redundant genome matching reads did not change. However, the complexity of the sample increased and became comparable to other samples with similar number of reads.

The next novel element was to evaluate the extent to which the data could be subsampled without affecting its structure we calculated the point-to-point PCC on expression levels of the original versus subsampled data from 95% to 40% of the original R set (Fig. S4). This showed that the correlations of abundantly expressed genes remained high over all subsamples, but that the correlation of low abundance transcripts decreased as the proportion of data subsampled dropped (note though that the variability between the original versus subsamples was lower than the variability between the biological replicates). We concluded that the subsampling was effective as it maintained high p2pPCC and strong concordance between the expression levels of the raw versus normalized data (Fig. S4).

The number of genes 'lost' due to the exclusion of some low abundance reads was typically <2%. Once we had determined that all samples passed the consistency check, we subsampled every sample to a fixed total of 50M reads and used to check whether subsamples were representative of the original data bootstrapping (Supplemental Material Methods 1). Following this step, one subsample was selected at random for each sample and used in the downstream analysis. This subsampling was efficient at correcting wide variation in read number, complexity differences and minimising the impact of normalization on the original data structure (Fig. 1).

The analysis of the distributions of complexity differences between replicates, coupled with the Jaccard similarity analyses applied on the normalized data, was used to identify outlier replicates, which were excluded from subsequent analyses. We classify as an outlier, samples for which the between-replicate similarity was higher than between-sample similarity, as shown by the Jaccard, complexity and p2pPCC analyses. In the 02RH example presented in Fig 1 we showed that the subsampling without replacement correctly highlighted replicate 3 as an outlier, while the subsampling with replacement did not. Following this post-normalization QC, we retained two biological replicates for each treatment for the *D. melanogaster* data. In general, we advocate the use of as many biological replicates as possible. However, as we show below, the analysis of subtle gene expression even with a limited number of replicates is possible and can be validated.

Subsampling with replacement vs subsampling without replacement

A subsampling with replacement normalisation has previously been proposed (Li, Tibshirani 2013b; Robinson, Storey 2014; Gierlinski et al. 2015); a version of the subsampling without replacement was proposed in (Stupnikov et al. 2015). The clear differences, and advantages, of a subsampling based method i.e. the lack of consistency of a scaling factor for all abundances, and thus the advantage over scaling methods, are discussed in (Li, Tibshirani 2013b; Stupnikov et al. 2015).

The main difference between the two approaches (with- vs without- replacement) is that for the former, the selection probabilities for the reads remains unchanged during the process, whereas for the latter the probabilities of selection are not constant, facilitating the selection of both high and low abundance reads. The abundance range for the later selection is wider and simulates the selection which takes place during the sequencing. The advantage of the former approach is that a higher number of reads can be achieved for a given sample (Li, Tibshirani 2013b) e.g. if a sample with 20M reads needs to be compared with a sample with 40M reads, then both samples could be subsampled, with replacement, at 30M reads. The down-side of the approach is that depending on the proportion of selected reads, over-amplification of high abundance reads and exclusion of low abundance reads can occur. This has a knock-on effect in amplifying the expression of abundant fragments and, as a consequence, reducing the expression of genes with low abundance which have fewer, low abundance incident fragments. In addition, the omission of low abundance reads may change the expression profile across transcripts (Fig 2(B)). The disadvantage of the subsampling without replacement is that it has an upper bound, calculated as the minimum sequencing depth between the samples of an experiment. Its advantage is that it renders the samples comparable (as described in the previous section) and allows the identification of potential outliers. In Fig 2(A), we compared the distribution of abundances for the two approaches (with versus without), on the same samples, to the same total. In Fig 2(B), we show the presence plots of a gene with 0.4 difference in complexity between the two approaches. The change in profile is visible for the third replicate, changes in the profile itself are highlighted with arrows.

Calling of DE using a hierarchical approach

To incorporate the experimental design into the DE call (Anders et al. 2010) we used a simple hierarchical approach for the prediction of DE transcripts (Supplemental Methods 2). The order of levels in the hierarchy was determined based on the amplitude of DE for each factor in the experiment (Fig. S5). For the *D. melanogaster* dataset the highest level in the hierarchy (i.e. that showed most DE) was body part (HT vs A), the second was ± rivals treatment. The distribution of DE between treatments and between replicates overlapped (Figure S5(A,B)), which indicated that the treatment DE was likely to be subtle.

We observed direct evidence of the biasing effect of low abundance FC (Fig. 3(A)). For example, using standard FC numerous low abundance DE genes in the HT were in fact a signature from the A body part (e.g. sperm and semen genes are specific to the A body part, but detected as differentially expressed in the HT at low abundance; Fig. 3(A,B)). The RNA-seq technique is highly sensitive and detects these transcripts due to leak through or movement of mRNAs. Unless a correction is applied, the list of DE genes is likely to contain numerous spurious and low-abundance entries. A practical solution was to use offset fold change (OFC), with the offset determined empirically from the data, as described in an earlier section, in preference to FC (Fig. 3A versus B) and to apply the hierarchical DE (Fig. 3C; Supplemental Methods 2). A comparison of the MA plots for all genes versus the A- and the HT-specific genes showed the effect of the hierarchical differential expression (Fig. 3C).

Case study - comparisons with existing approaches

We next evaluated the output gained from the analysis of the *D. melanogaster* data using our bioinformatics framework, with that obtained from the analysis of the same, original input data (consisting of all 3 replicates for each condition) using DESeq2 (Love et al. 2014a) and edgeR (Robinson, Oshlack 2010; Zhou et al. 2014). Although the DE was conducted on the HT samples, both the HT and AB samples were given as input the edgeR and DEseq2 analyses.

Effect of the normalization

An *a priori* (and necessary, but not sufficient) condition for reliable DE call is good comparability between the distributions of expression levels. To assess whether the normalized expression levels became more similar after the normalization (in particular, whether the replicates became comparable), we compared the distributions of expression levels (log_2 scale) for the raw data versus RPM, quantile, subsampling with and without replacement, DEseq2 and edgeR normalizations (Fig. 4). The boxplot of the raw abundances (Fig. 4A) illustrates the variation among the replicates and samples and clearly indicated that normalization was required. The RPM normalization (Fig. 4B) rendered the A and HT distributions comparable to some extent. However, it was difficult to separate the A- or HT-specific genes. In addition, variability between samples was still present (especially for the HT samples). The quantile normalization (Fig. 4C) rendered the distributions comparable, as did the subsampling with and without replacement (Fig. 4D1 and Fig 4D2). DEseq2 performed well (Fig. 4E) – although residual differences in the distributions of the A vs HT samples remained. EdgeR (Fig. 4F) did not effectively equalize the distributions of abundances. We concluded that the subsampling, quantile and DESeq2 normalizations (Fig. 4C,D1, D2,E) were most effective at producing comparable distributions of normalized expression levels for this dataset.

Differences in the DE call between methods

To evaluate the effect of the normalization and hierarchical DE call we compared analyses of the *D. melanogaster* 2h HT and A samples ±rivals (using as input all three biological replicates of the original data) with the output of DEseq2 and edgeR (Fig. 5).

The subsampling without replacement normalization and hierarchical DE call (Fig. 5A) showed a relatively low number of up/down-regulated genes with relevant biological functions. The equivalent analysis for DEseq2 (Fig. 5*B*) called many more genes as DE that fell in the region of +/- 0.5 log_2 FC (i.e. below the validatable threshold (Morey et al. 2006) (Fig. S5). The analysis using edgeR (Fig. 5C) showed a high frequency of low abundance DE and of leaky genes, which is likely to either represent noise or biological signal of an insufficient magnitude to be captured effectively in the low throughput validation (Fig. S5). The degree of overlap between the three methods (Fig. 5D) revealed a small number of core genes present in the intersection. edgeR and DEseq2 called many more genes as DE and the number of genes uniquely identified by edgeR and DEseq2 was also larger than the number identified in common between the two. These results show that the pipeline chosen will have a strong effect on the biological interpretation of the DE analysis.

For the ±rivals comparison for the HT samples, out of the 575 genes that were specific to edgeR, 14 were HT genes (all with max abundance > 50) and 561 were A genes (327 with max abundance > 50 and 234 < 50; Fig. S5). Out of the 578 genes specific to DESeq2, 101 were HT genes (100 with abundance > 50) and 477 were A genes (271 with max abundance > 50 and 206 < 50; Fig. S5). The predominance of A genes in the DE call supported the use of the hierarchical DE approach. The presence of low abundance genes supported the use of an offset for the calculation of DE.

Of some concern was that for genes with a reasonable abundance (> 50) the expression intervals for the ± rivals differences called by DEseq2 and edgeR were close/overlapping, making independent validation using low throughput methods challenging. These results showed evidence of a high number of DE genes called by both or either of DEseq2 and edgeR that would be difficult to validate independently.

Reasons for the differences in DE call between methods are likely to result from a propagation of factors throughout the whole analysis. These potentially include low comparability of normalized distributions of expression or the imperfect assessment of hierarchical distribution of DE levels in the experiment. In the calculation of DE by DEseq2, replicate-to-replicate variability is averaged and DE over and above this variation is calculated. The accuracy of such averaging relies on replicates having a low coefficient of variation (CV = standard deviation/mean). However, in the *D. melanogaster* data, in many cases the CV was over 0.25 and in some was > 0.5 (across all abundances, Fig. S6). In the example shown (Figs. S6A-D), there was clearly higher dispersion (CV) at low transcript abundance and consistently high CV across higher abundances. This variation was corrected in our analyses by the subsampling normalization and the use of an OFC (Fig. S6 E,F,G,H, in which dispersion showed minimal variation across transcript abundance and the CV was consistently low (generally < 0.1)). The newer version of DESeq, DESeq2, (Love et al. 2014a) notes the effect of high replicate variation and proposes a shrinkage estimation for dispersions based on empirical Bayes and FC, to improve stability and interpretability of estimates. Our results showed that these changes helped to minimize, but did not fully solve, the overall issue of high variability in transcript abundance.

Comparison of low throughput validated genes with DESeq2 and edgeR outputs

We next investigated whether the set of DE genes identified using the hierarchical approach from our *D. melanogaster* dataset, and validated by qRT-PCR (Mohorianu et al. 2017), were present in the output of DEseq2 and edgeR. Reassuringly, based on DESeq2 our qRT-PCR reference genes were not called DE. Two other genes of interest from the A samples that were validated as DE, had a p-value < 0.05 by DESeq2 (although adjusted p-value > 0.05). One gene of interest validated from the HT had a p-value < 0.05 (but again not according to the adjusted p-value). For DESeq2 the $\log_2$(FC) values were small (0.15 and 0.16, respectively). Hence these genes were not likely to have been selected for further investigation. Based on the edgeR output, our reference genes were also determined as not significantly DE. For the validated A genes of interest (GOI), only one was called marginally DE by edgeR (p = 0.08, FBgn0259998, but with small $\log_2$(FC) = 0.24). For the HT, one GOI, with $\log_2$(FC) = 0.58, was called as significantly DE (the same gene as identified by DESeq2). Another GOI, FBgn0044812, was identified with $\log_2$(FC) = 0.82 yet with a p-value of 0.49 from edgeR and therefore would not have been selected (Table S1).

Comparing the validated gene set with the output of edgeR and DESeq2, we conclude that some GOIs failed to be identified and therefore the corresponding biological functions (immunity, odorant perception) might have been overlooked.

Which normalization to choose?

An RNA-seq sample is a snapshot of RNA fragments present at a given time, randomly selected according to the RNA abundances, to fill the sequencing space. Due to the stochastic nature of the sequencing process, even technical replicates, at different sequencing depths, do not exhibit a constant scaling factor for all abundances. Also, RNA-seq outputs have varying fits to standard distributions, making it difficult to define "the best" choice. Although the subsampling without replacement normalization was efficient in minimizing the effects of the variable sequencing depth, while preserving a high similarity with the original samples (Fig. 3), we suggest that it is advisable to test different normalizations on mRNA-seq and choose the most appropriate method for the given dataset, on a case-by-case basis (Beckers et al. 2017)

Analysis of human mRNA-seq datasets using subsampling (without replacement) normalization

RNA-seq is expected to have good external validity and produce comparable results when the same RNA is used, across different laboratories. However, a recent study of mRNA-seq conducted on the same human samples (expression level variation in lymphblastoid cell lines) involved the use of the same samples sequenced in two different locations: Yale versus Argonne (Pickrell et al. 2010). Some variation between the results from the different laboratories was observed (Zhou et al. 2014). The authors analysed these data further to explore whether edgeR could reduce the variability between replicates. We tested whether our subsampling normalization could further reduce such variation. To do this, we randomly selected 5 sets of samples (144, 153, 201, 209 and 210) with two replicates each, one from the Yale laboratory source and one from the Argonne source. For these runs, the length of the reads was 36nt for Yale and 46nt for the Argonne-derived data. Since the length of the sequencing read influences the number of unique fragments and the mapping to the reference transcriptome (and, as a result, the gene expression) we trimmed all reads to comparable lengths (35nt) and mapped the reads to the reference human genome using full length, no mis-match or gap criteria and using PatMaN (Prufer et al. 2008). The subsampling was conducted on 7M reads (the number of reads for the smallest sample was 7.1M, and for the largest sample was 8.7M)

We created comparable plots to (Zhou et al. 2014) for the data subjected to subsampling normalization. MA plots for the two replicates of each sample showed high reproducibility between runs. The distribution of the coefficient of variation (CV) versus the abundance for the 5 selected sets of samples with two replicates each (one for Yale and one for Argonne) (Fig. S7) showed that the CV for all 5 pairs of samples was consistently (< 0.1) lower than for the analysis of (Zhou et al. 2014)) indicating a very high similarity between the runs. The MA plots on the same sets of two samples showed a high reproducibility between replicates (no genes showing $|\log_2(OFC)| > 1$). The genes showing DE were mainly localized in the 2^4 (16) – 2^6 (64) range, which is borderline for validation/noise. Together, these analyses showed that: (i) the CV obtained when the subsampling (without replacement) normalization was employed was lower than the CV reported in (Zhou et al. 2014), suggesting that the normalization was tighter, (ii) there was very little DE between replicates, indicating good reproducibility between the sequencing runs.

Overall, we conclude that the subsampling, without replacement approach cleared the technical differences between the two runs in the different laboratories and this approach rendered the samples comparable, potentially improving the biological inference.

Conclusion

The main conclusion from this study was to emphasise the need to check multiple approaches for the analysis of a dataset and to show that both qualitative and quantitative QC are informative, and the applicability of subsampling (without replacement) -based normalization and hierarchical structuring of the DE call, is efficient in managing variation in read number and differences in sample complexities. In comparison to existing methods, the adapted methods performed well and identified valid candidates that were confirmed using low throughput approaches (Mohorianu et al. 2017). We also successfully applied the subsampling (without replacement) normalization to existing mRNA-seq datasets, used to analyse inter-laboratory variation (Pickrell et al. 2010); the adapted approach proved to be efficient in comparison with existing methods at minimizing potentially confounding sources of variation. Determination of accurate gene expression levels is essential for all mRNA profiles but is also key to successful correlation analysis between mRNAs and sRNAs (Mohorianu et al. 2012; Mohorianu et al. 2013).

METHODS

Quality check (QC)

For the mRNA-seq samples, the QC consisted of two stages. Stage 1 comprised of previously described methods (Conesa et al. 2016a) including: (i) the analysis of FastQ quality scores (Andrews 2010), (ii) the total number of reads (sequencing depth) and the read duplication rate, defined as complexity (Mohorianu et al. 2011a), (iii) nucleotide composition relative to the genome and transcriptome of *D. melanogaster*, used to highlight biases such as PCR and ligation bias (Sorefan et al. 2012), (iv) strand bias quantified on CDS incident reads as $|P - 0.5| + |N - 0.5|$, where P and N were the proportion of positive and negative strand read matches, respectively (Mohorianu et al. 2011a) and (v) proportions of reads matching the different genome annotation classes (e.g. mRNAs, t/rRNAs, miRNAs, UTRs, introns, intergenic regions (Conesa et al. 2016a); matching was done on full length reads with no mis-matches or gaps allowed, using PatMaN, (Prufer et al. 2008)). Stage 2 comprised of quantitative approaches, some applied/designed on mRNA-seq data for the first time, which provided an increased insight into sample comparability and enabled us to evaluate the effectiveness of the normalization. The expression level of a gene/ transcript was calculated as the algebraic sum of the raw/normalized abundances of the incident reads (Mortazavi et al. 2008a). We examined (i) sample similarity calculated using the Jaccard similarity index (Jaccard 1901) on the top 1000 most abundant genes, and intersection analyses); theseq measures were calculated as the ratio between number of genes found in common to the number of unique genes present in in either samples, (ii) complexities (calculated at gene level and presented as Bland-Altman plots) and (iii) point-to-point PCC between gene expression profiles in different replicates/ samples. The latter were computed on the vector of expression defined for each gene. For all positions *i* on a gene we computed y[i] which is the sum of abundances of fragments incident with position *i*. The point-to-point PCC was computed as the standard PCC on the corresponding vectors from the two samples which were compared.

Normalization

We adapted a normalization procedure based on subsampling (without replacement) (Li et al. 2013a); the consistency of the subsample was validated using bootstrapping. The subsampling, without replacement, was done on the redundant set of reads (before genome matching, with the ncRNAs incident reads removed). The proportion of genome matching reads and the variation in gene complexities (coupled with the p2pPCC between the subsamples and the original sample) were used as criteria for consistency of the subsamples. Each sample was first subjected to incremental subsampling in order to investigate the effect on the data structure (complexities, both for non-matching and genome-matching reads) of sampling 95% through to 45% of the data, with successive decreasing steps of 5%. A sample was deemed satisfactory if the proportion of redundant genome matching reads remained constant and the average point-to-point PCC were above 95% as the number of redundant reads was decreased from 95% to 45%. This step represented an empirical determination of the level of subsampling that could be done whilst preserving the original data structure. The second step of the normalization was the subsampling to a fixed total (the minimum sequencing depth of the accepted samples). Samples with low sequencing depths, which would lead to a heavy subsampling for the samples with high read numbers (less than 55%, empirically determined), were treated on a case-by-case basis. A quantile normalization (Bolstad et al. 2003) may be employed after this step to render the distributions fully comparable. The pseudocode is presented in Supplemental Methods 1.

Existing procedures which were used for the comparison of the new normalization methods were: scaling normalization (Mortazavi et al. 2008a), for which the scaling total was the mean of the sequencing depths of the compared samples, quantile normalization (Bolstad et al. 2003) and the normalization approaches from edgeR (Zhou et al. 2014) and DESeq2 (Love et al. 2014a). All were employed using the recommended standard parameters.

Differential Expression call

Existing methods for the DE call are often based on comparing the variability between replicates with the difference between the treatments. However, calculation of variance (or cv) based on a small number of points may often not reflect the true variance of the given gene/transcript (Krzywinski et al. 2013; Blainey et al. 2014; Altman et al. 2015). Moreover, when small numbers of measurements are available, a more conservative approach, which we use here, is to approximate that replicate measurements will fall within the two limits of the maximal interval (Claridge-Chang et al. 2016).

The maximal confidence intervals are defined on the minimum and maximum normalized expression levels for the replicated measurements. The amplitude of the DE is calculated on a worst-case scenario, on the proximal ends of the maximal intervals i.e. this method ensures that all points in the treatment are on one side (for up-regulation, above and for down-regulation, below) of the control measurements (Beckers et al. 2017; Collins et al. 2017). As a result of the stringency of this approach all genes called DE using these rules will also be called DE under all statistical tests. In addition, the threshold on the amplitude of the DE (for (Mohorianu et al. 2017) set at 1.5 fold change, in line with the empirical threshold described in Morey et al) prevents the selection of genes with separate but close expression ranges and ensures a higher chance for validation confirmations.

DE was calculated using a hierarchical approach and by applying an offset fold change (OFC) method (with offset=20, empirically determined, using the point-to-point PCC, for all replicates within all samples). There were 3 steps to the hierarchical analysis used for the analysis of the *D. melanogaster* transcriptome data. (i) identification of levels for the hierarchical differential expression and the constituent internal classes. For the *D. melanogaster* data one 'level' was body part (with HT and A as internal classes) and the other was treatment (with presence or absence of rivals as classes). (ii) the ordering of the hierarchical levels based on the amplitude of differential expression. This was quantified by the width/ spread of the distribution of DE in terms of mean/ median, IQR and min/max values. The amplitude of DE in descending order provided the correct ordering of the levels for the hierarchical DE. (iii) the DE analysis on the proximal ends of the CIs, using OFC (Mohorianu et al. 2011a). The pseudocode is presented in Supplemental Methods 2.

The two-step DE procedure (using OFC) consisted of (i) calculation of the list of genes showing DE between body parts, followed by (ii) calculation of the DE between genes in the ± rivals treatment comparisons. Step (i) was conducted on the summed expression levels in the ± rivals pairs (i.e. the ± HT samples combined, and the ± A samples combined, for all time points). The genes were then separated into genes expressed only in HT, only in A, and in both the HT and A. Step (ii) of the DE was then applied on the resulting 3 categories (HT; A; HT+A) using the ± rival condition. We called DE the genes which showed after the second step of the DE described above, of more than 1.5 fold between the treatments (+/- rivals). The DE call as determined by edgeR and DESeq2 were calculated using the default functions and parameters.

DATA ACCESS

mRNA samples: (a) *D melanogaster*: males of *D melanogaster* exposed to conspecific rivals (or not) for 3 time periods (GSE55930). (b) *H sapiens*: For the mRNA Human samples, we chose 5 samples from the Pickrell et al. 2010 (Pickrell et al. 2010) study (GSE19480) in order to compare gene expression variation in RNA sequencing between the Argonne and the Yale laboratory sequencing runs. The selected samples were: GSM485369 (NA19144_yale), GSM485380 (NA19144_argonne); GSM485368 (NA19153_yale), GSM485383 (NA19153_argonne); GSM485367 (NA19201_yale), GSM485381 (NA19201_argonne);

GSM485365 (NA19209_yale), GSM485388 (NA19209_argonne); GSM485364 (NA19210_yale), GSM485382 (NA19210_argonne). These samples were derived from lymphoblastoid cell lines (LCLs) derived from unrelated individuals from Nigeria (extensively genotyped by the International HapMap Project). The sequencing was done on Illumina GAII, with sequencing reads of 36nt, for the Yale sequencing samples and 46nt for the Argonne sequencing.

ACKNOWLEDGEMENTS

I would like to thank Amanda Bretman, Damian T. Smith, Emily Fowler and Tracey Chapman for generating the *D. melanogaster* data and for enriching discussions during the preparation of this manuscript. I am also grateful to Rachel Rusholme Pilcher and Anders Wiren for helpful suggestions during the writing of this manuscript. The work was funded by the BBSRC and the NERC (research grants BB/H002499/1, BB/L003139/1 and NE/J024244/1) to TC.

COMPETING INTERESTS

The authors declare there are no competing interests.

SUPPLEMENTAL MATERIAL

SUPPLEMENTAL METHODS

SUPPLEMENTAL METHODS 1 - Subsampling normalization – pseudocode. A description with details for (1) Incremental subsampling and bootstrapping check for consistency of a sample, and (2) Subsampling to a fixed total.

SUPPLEMENTAL METHODS 2 - Two step (Hierarchical) differential expression (HDE) - pseudocode. A description with technical details for the two step (hierarchical) DE, including the identification of levels in the hierarchy.

SUPPLEMENTAL TABLES

TABLE S1. Annotation overview for the 02 samples in the *D. melanogaster* dataset. For each 02 samples (described in Mohorianu et al 2017) we present the number and proportions of reads, matching to the *D. melanogaster* genome (v 6.11) and to the corresponding annotations (exons, introns, 5' and 3' UTRs, ncRNAs and intergenic regions)

TABLE S2. Example of intersection analysis for the 02-A, 02+A, 02-H and 02+H samples in the *D. melanogaster* dataset. Replicates 1 samples 02-A, 02+A, 02-H and 02+H were used to illustrate the proportion of reads mapping simultaneously to pairwise groups of CDSs, exons, 5’ and 3’ UTRs, introns and intergenic regions. We observed a high proportion of exon matching reads present on 3’ and 5’ UTR. In the main study we computed expression levels using gene mapping reads.

TABLE S3. Jaccard similarity index on the 02 samples in the *D. melanogaster* dataset The Jaccard similarity at gene level was computed on the top 1000 most abundant genes in each sample (out of a total of 15 513 genes expressed in at least one sample). As a result, it is not biased by the different number of genes present in each sample. Shown is a 12 by 12 matrix of all the original samples compared with each other. Samples are labelled by time point (2h), by ± rivals treatment, by body part (A or HT) and then by replicate number. Each

sample tested against itself along the diagonal is therefore 100% similar and shares the top 1000 most abundant genes in common. A to A comparisons are shaded in purple, HT to HT comparisons in peach. Samples drawn from the same body parts shared > 90% similarity, and between body parts the similarity dropped to ~50%. Similarity between the ± rivals treatments tended to be higher than between replicates. Two illustrative examples are highlighted, in which ± rivals indices (in red bold) were generally higher than replicate to replicate similarity (blue bold). This highlighted the need for the adapted normalization methods.

TABLE S4. Example of incremental check for subsampling without replacement for sample 02EH2 in the *D. melanogaster* dataset.

For sample 02EH2, we present the incremental subsampling, without replacement. To judge whether a sample is consistent, and to determine the consistency threshold, we use the proportion of redundant reads matching to the referenece genome. As a consequence of the incremental subsampling, the complexity increases. A replicate is accepted if it exhibits a similar complexity (and distribution of per-gene complexities) with the other replicates of the same type of sample.

TABLE S5. Results from (A) DEseq2 and (B) edgeR analyses of the *Drosophila melanogaster* qRT-PCR 'validated' gene set. For the validations we used 3 reference genes and validated 15 A genes and 6 HT genes based on the DE selection using subsampling normalization and hierarchical DE. We investigated whether these genes were called DE by either (A) DESeq2 or (B) edgeR. In Table 1A we present the results for DESeq2, in Table 1B the results for edgeR. For each of the three categories of genes (reference genes, AB genes and HT genes) we show the average of normalized abundances (baseMean for DESeq2 and logCPM for edgeR), the fold change between treatments (log2 FoldChange for DESeq2 and log2FC for edgeR) and the DE p-value and adjusted p-value (used for the DE call).

SUPPLEMENTAL FIGURES

FIGURE S1 Analysis framework for mRNA-seq data.

Inputs are shown (sequencing data in FASTQ format, and the corresponding reference genome and transcriptome in FASTA/GFF) and the six main steps: Quality check (QC), alignment, normalization of gene abundances, identification of DE, functional enrichment and finally low-throughput validation.

FIGURE S2. Correlation analyses (Pearson, Spearman and Kendall correlation coefficients) between the gene expression levels for the *D. melanogaster* data for (A) all samples, (B) HT samples, (C) A samples. A1, B1, C1 show the PCC; A2, B2, C2 show the SCC; A3, B3 and C3 show the KCC. Each panel shows the distributions of correlation coefficients for all pairwise comparisons. For example, in panel A.1, sample 1 on the x-axis shows the distribution of the n=35 correlation coefficients calculated between the gene expressions in sample 1 compared with gene expressions in all other 35 samples using PCC. The results are presented as a standard boxplots.

FIGURE S3. Distribution of point-to-point PCC between gene expression profiles against gene expression levels ($\log_2$ scale) for pairwise comparisons for the *D. melanogaster* data for the 3 replicates of the 02HT- sample as an example (2h, HT body part, no rivals). Panel a shows replicate 1 vs 2, b replicate 1 vs 3 and C replicate 2 vs 3. Shown are the raw data, prior to normalization. For all replicate comparisons, more variability is consistently observed at lower abundances.

FIGURE S4. Point-to-point PCC between the raw and subsampled data of the *D. melanogaster* data. To show the consistency during the subsampling, shown are the point-to-point PCC between the original data and the data incrementally subsampled from 40% to 95% (Panels A to L). On the x-axis is the gene abundance ($\log_2$) and on the y-axis the distribution of point-to-point PCCs calculated for each expressed gene.

FIGURE S5. Identification of the hierarchy levels for the hierarchical differential expression (HDE) analysis based on the distribution of DE for the different classes of samples, i.e. replicates, body parts and ± rivals treatments (for the *D. melanogaster* data).

Frequency plots were used to show the distribution of DE between samples. Panel A shows the replicate-replicate DE (blue) and the with/without rivals DE (red) for the abdomen (A) samples. Panel B shows the corresponding data for the HT body part. Panel C shows the distribution of DE for the with/without rivals treatments (blue for HT and green for A samples) and the DE between HT and A (orange).

FIGURE S6. **Distribution of abundances for the *D. melanogaster* data (for the ± rivals treatment DE) for genes identified as DE exclusively by each method**. EdgeR only genes are presented in 5A, DEseq2 only genes in 5B and subsampling normalization only genes in 5C. For each gene (FBgn identifier) identified as DE exclusively by each method, the normalized abundance is given for each of the 2h HT and A ± rivals samples. The predominance of leaky genes in the DE calls of edgeR and DESeq2 highlighted the need for the hierarchical DE. The presence of low abundance genes indicated the requirement for an offset for the calculation of the extent of DE.

FIGURE S7. **Comparison of the coefficient of variation applied on the *D. melanogaster* data**. On the x-axis is the abundance in $\log_2$ scale, on the y-axis we represent the coefficient of variation (cv), defined as the ratio between the standard deviation and the mean. For clarity, the distributions are represented as standard boxplots. The upper panels (A,B,C,D) show the cv for the original data for A samples, without rivals, A samples with rivals, HT samples without rivals and HT samples with rivals, respectively. The lower panels (E,F,G,H) give the CV for the same samples, after the subsampling normalization. The horizontal lines indicate 0.5 and 0,25 cv, to ease visualization. It is clear that the subsampling normalization reduced the variance between the replicates to < 0.25 cv across most abundances (Panels E-H), whereas the cv was much higher across all abundances for the raw data (Panels A-D).

FIGURE S8. **Analysis of the effect of the subsampling normalization on technical (laboratory-laboratory) variation in mRNA-seq for the human mRNA-seq data in (Pickrell et al. 2010)**. In the upper plots we show the coefficient of variation (CV) obtained after the subsampling normalization for 5 sequencing pairs (each pair consisted of a Yale laboratory run compared to an Argonne run: A1,A2 = Sample 144; B1,B2 = Sample 153; C1,C2 = Sample 201; D1,D2 = Sample 209; E1,E2 = Sample 210). Shown is the CV against abundance ($\log_2$ scale). For all comparisons we achieved lower CVs in comparison to the Zhou et al (2014) analysis of these sample data. In red we represent the CV of these data obtained using edgeR, in blue the CV using DESeq2. It is evident that our subsampling normalization achieved lower CV across all abundances in comparison to both edgeR and DRseq2. Based on these distributions we conclude that the samples from the different laboratories can be rendered comparable using the subsampling approach, i.e. the subsampling normalization removed the technical differences between the two different laboratory runs. In the lower panels, we present the MA plots, after the subsampling normalization, for the same pairs of samples. The tightness of these plots (all falling within ±0.5 OFC) supports the conclusion that the subsampling has rendered these samples derived from sequencing in different laboratories highly comparable.

REFERENCES

Aanes H, Winata C, Moen LF, Ostrup O, Mathavan S, Collas P, Rognes T, Alestrom P. 2014. Normalization of RNA-sequencing data from samples with varying mRNA levels. *PLoS One* 9:e89158.

Altman N, Krzywinski M. 2015. Points of significance: Sources of variation. *Nat Methods* 12:5-6.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* 11:R106.

Andrews S. 2010. FastQC: A quality control tool for high throughput sequence data. *abraham Bioinformatics*.

Beckers ML, Mohorianu I, Stocks MB, Applegate C, Dalmay T, Moulton V. 2017. Comprehensive processing of high throughput small RNA sequencing data including quality checking, normalization and differential expression analysis using the UEA sRNA Workbench. *RNA*.

Blainey P, Krzywinski M, Altman N. 2014. Points of significance: replication. *Nat Methods* 11:879-880.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19:185-193.

Bullard JH, Purdom E, Hansen KD, Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94.

Claridge-Chang A, Assam PN. 2016. Estimation statistics should replace significance testing. *Nat Methods* 13:108-109.

Collins DH, Mohorianu I, Beckers M, Moulton V, Dalmay T, Bourke AF. 2017. MicroRNAs Associated with Caste Determination and Differentiation in a Primitively Eusocial Insect. *Sci Rep* 7:45674.

Conesa A, Madrigal P, Tarazona S, et al. 2016a. A survey of best practices for RNA-seq data analysis. *Genome Biol* 17:13.

Conesa A, Madrigal P, Tarazona S, et al. 2016b. A survey of best practices for RNA-seq data analysis. *Genome biology* 17:13.

Consortium SM-I. 2014. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol* 32:903-914.

Cui X, Churchill GA. 2003. Statistical tests for differential expression in cDNA microarray experiments. *Genome biology* 4:210.

DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G. 2012. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* 28:1530-1532.

Dillies MA, Rau A, Aubert J, et al. 2013a. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics* 14:671-683.

Dillies MA, Rau A, Aubert J, et al. 2013b. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 14:671-683.

Evans C, Hardin J, Stoebel DM. 2017. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinform*.

Fang Z, Cui X. 2011. Design and validation issues in RNA-seq experiments. *Brief Bioinform* 12:280-287.

Gierlinski M, Cole C, Schofield P, et al. 2015. Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics* 31:3625-3630.

Hansen KD, Brenner SE, Dudoit S. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research* 38:e131.

Jaccard P. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*:547-579.

Jackson TJ, Spriggs RV, Burgoyne NJ, Jones C, Willis AE. 2014. Evaluating bias-reducing protocols for RNA sequencing library preparation. *BMC Genomics* 15:569.

Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R. 2011a. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res* 39:e141.

Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R. 2011b. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic acids research* 39:e141.

Jia B, Xu S, Xiao G, Lamba V, Liang F. 2017. Learning gene regulatory networks from next generation sequencing data. *Biometrics*.

Krzywinski M, Altman N. 2013. Significance, P values and t-tests. *Nat Methods* 10:1041-1042.

Li J, Tibshirani R. 2013a. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical methods in medical research* 22:519-536.

Li J, Tibshirani R. 2013b. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat Methods Med Res* 22:519-536.

Li P, Piao Y, Shon HS, Ryu KH. 2015. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics* 16:347.

Li S, Labaj PP, Zumbo P, et al. 2014. Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat Biotechnol* 32:888-895.

Lin Y, Golovnina K, Chen ZX, Lee HN, Negron YL, Sultana H, Oliver B, Harbison ST. 2016. Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster. *BMC Genomics* 17:28.

Love MI, Huber W, Anders S. 2014a. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* 15:550.

Love MI, Huber W, Anders S. 2014b. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15:550.

McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ, Nuzhdin SV. 2011. RNA-seq: technical variability and sampling. *BMC Genomics* 12:293.

Mohorianu, II, Bretman A, Smith DT, Fowler E, Dalmay T, Chapman T. 2017. Genomic responses to socio-sexual environment in male Drosophila melanogaster exposed to conspecific rivals. *RNA*.

Mohorianu I, Lopez-Gomollon S, Schwach F, Dalmay T, Moulton V. 2012. FiRePat - Finding Regulatory Patterns between sRNAs and Genes. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery* 2:273-284.

Mohorianu I, Schwach F, Jing R, Lopez-Gomollon S, Moxon S, Szittya G, Sorefan K, Moulton V, Dalmay T. 2011a. Profiling of short RNAs during fleshy fruit development reveals stage-specific sRNAome expression patterns. *The Plant journal : for cell and molecular biology* 67:232-246.

Mohorianu I, Schwach F, Jing R, Lopez-Gomollon S, Moxon S, Szittya G, Sorefan K, Moulton V, Dalmay T. 2011b. Profiling of short RNAs during fleshy fruit development reveals stage-specific sRNAome expression patterns. *Plant J* 67:232-246.

Mohorianu I, Stocks MB, Wood J, Dalmay T, Moulton V. 2013. CoLIde: a bioinformatics tool for CO-expression-based small RNA Loci Identification using high-throughput sequencing data. *RNA biology* 10:1221-1230.

Morey JS, Ryan JC, Van Dolah FM. 2006. Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. *Biological procedures online* 8:175-193.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008a. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* 5:621-628.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008b. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621-628.

Ozsolak F, Milos PM. 2011. RNA sequencing: advances, challenges and opportunities. *Nature reviews. Genetics* 12:87-98.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 14:417-419.

Pickrell JK, Marioni JC, Pai AA, et al. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464:768-772.

Prufer K, Stenzel U, Dannemann M, Green RE, Lachmann M, Kelso J. 2008. PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics* 24:1530-1531.

Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D. 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* 14:R95.

Risso D, Schwartz K, Sherlock G, Dudoit S. 2011. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* 12:480.

Robinson DG, Storey JD. 2014. subSeq: determining appropriate sequencing depth through efficient read subsampling. *Bioinformatics* 30:3424-3426.

Robinson DG, Wang JY, Storey JD. 2015. A nested parallel experiment demonstrates differences in intensity-dependence between RNA-seq and microarrays. *Nucleic Acids Res* 43:e131.

Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology* 11:R25.

Roca CP, Gomes SI, Amorim MJ, Scott-Fordsmand JJ. 2017. Variation-preserving normalization unveils blind spots in gene expression profiling. *Sci Rep* 7:42460.

Schmieder R, Lim YW, Rohwer F, Edwards R. 2010. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC bioinformatics* 11:341.

Schurch NJ, Schofield P, Gierlinski M, et al. 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* 22:839-851.

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

Sorefan K, Pais H, Hall AE, Kozomara A, Griffiths-Jones S, Moulton V, Dalmay T. 2012. Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence* 3:4.

Stupnikov A, Glazko GV, Emmert-Streib F. 2015. Effects of subsampling on characteristics of RNA-seq data from triple-negative breast cancer patients. *Chin J Cancer* 34:427-438.

Torres-Oliva M, Almudi I, McGregor AP, Posnien N. 2016. A robust (re-)annotation approach to generate unbiased mapping references for RNA-seq-based analyses of differential expression across closely related species. *BMC Genomics* 17:392.

Trapnell C, Roberts A, Goff L, et al. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* 7:562-578.

Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* 131:281-285.

Wang L, Wang S, Li W. 2012. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 28:2184-2185.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* 10:57-63.

Zhou X, Lindsay H, Robinson MD. 2014. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic acids research* 42:e91.

FIGURES

FIGURE 1 Distributions of complexities, calculated at the gene level, on the *D. melanogaster* mRNA-seq data.

Transcript abundances (x-axis, $\log_2$ scale) are plotted against the absolute difference in complexities, i.e. the non-redundant/redundant (NR/R) ratio (y-axis) for all genes and all biological replicate comparisons (a = replicate 1 vs 2; b = replicate 1 vs 3; c = replicate 2 vs 3. Example data shown are for the three original replicates of the 02+H (2 hours, rivals present, head thorax) samples. The differences in complexities were calculated on the raw data (top row), and on the data after subsampling normalization with replacement (middle row) and without replacement (bottom row). Red horizontal lines indicate 0.05 and 0.1 differences in complexity. Before subsampling the complexity differences were frequently > 0.1. The subsampling approaches (with or without replacement) rendered the biological replicates more comparable and reduced the complexity differences to < 0.1 across all transcript abundances. In addition, the subsampling without replacement maintained the conclusion that the third replicate (R3) was problematic, whereas the subsampling with replacement masked this conclusion.

FIGURE 2 Comparison of results obtained using the subsampling with or without replacement. On the top row we present the MA plots on the gene expression levels, normalized using either the with- or without- replacement approaches, for the three replicates of the 02+H sample. Although the variability between the two approaches is contained within the +/- 0.5 $\log_2$(OFC), we observe a higher variability in expression for the low abundance genes. In subplot (B) we present the presence plots for the gene FBgn0033865 for each of the three replicates (the individual panels) obtained using either the subsampling without replacement (black solid line) or subsampling with replacement (red solid line). The arrows indicate the regions where the two approaches provide different answers. The arrow indicating the first exon of the gene highlights the difference observed for the third replicate (02+H, rep3).

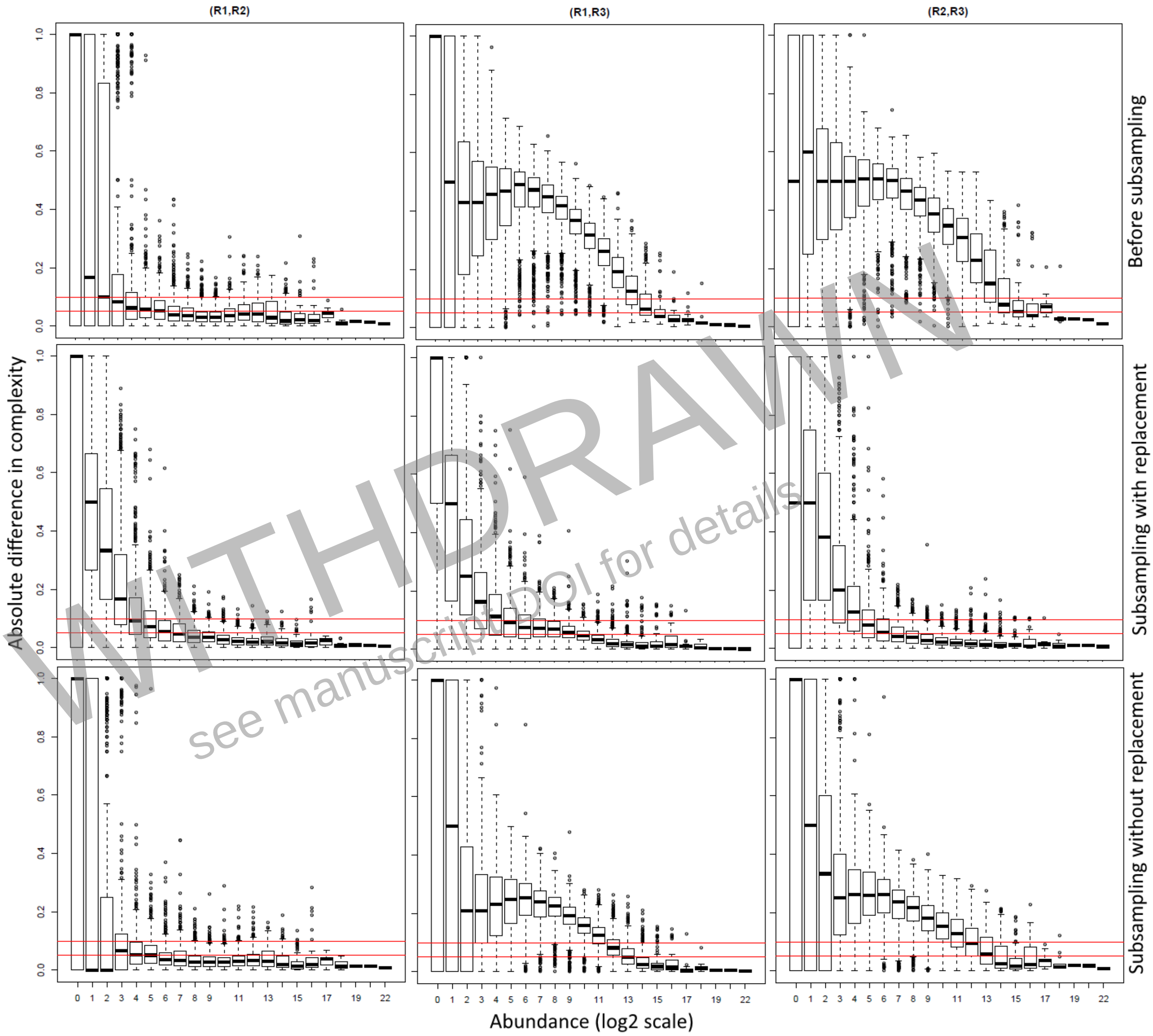
FIGURE 3 Distribution of DE as calculated by using fold change (FC) versus offset fold change (OFC) and the effect of incorporating hierarchical DE (HDE). Shown are MA plots (x-axis showing gene abundance ($\log_2$), y-axis indicating FC/OFC for replicate-to-replicate comparisons for the 2h samples. Panels A1, B1, C1 show 02A- comparisons, panels A2, B2, C2 for 02A+ samples, A3, B3, C3 for 02HT- and A4, B4, C4 for 02HT+ samples (Sample codes: 02 = 2h of exposure, A = abdomen, HT = head-thorax, + = with rivals, - = without rivals). Panel a shows the distribution of DE calculated using FC, showing how the low abundance genes distort the distribution of DE. Panel b shows the DE distribution using OFC (offset=20). Here most of the low abundance genes were excluded. Panel c shows the DE distribution following hierarchical DE analysis using OFC for A- and HT-specific genes highlighting the elimination of low abundance, potentially spurious, DE. The red horizontal lines denote 0 $\log_2$ FC/OFC and the blue lines ± 0.5 $\log_2$ FC/OFC.

FIGURE 4 Comparison of expression distributions resulting from different normalization methods. Shown are standard boxplots of normalized gene expressions. On the x-axis are the different samples (e.g. 02A-1 = 2h time point abdomen body part, no rivals, replicate 1) and the on the y-axis the $\log_2$ gene expression. Panel A shows the raw expression levels, B the RPM normalization to a fixed total of 50M reads, C the quantile normalization, D1 the subsampling (with replacement) normalization to a total of 50M reads D2, the subsampling without replacement, E the DESeq2 normalization and F the edgeR normalization. Effective normalization (e.g. C and D) is observed when the distributions become most comparable.

FIGURE 5 Comparison of distribution of DE obtained using the subsampling normalization and HDE, DESeq2 and edgeR. MA plots, with x-axis showing $\log_2$ average abundances against OFC with an offset of 20 (Panel A) and FC (Panels B and C). The example shown ism for the 02HT ± rivals DE comparison. The red line indicates 0 $\log_2$ FC/OFC and the blue lines ±0.5 $\log_2$ FC/OFC. Red data points represent the genes ‘called’ differentially expressed by each of the methods. Panel A shows the results for subsampling normalization

with DE calculated using the hierarchical approach, Panel B for DEseq2 and Panel C for edgeR. Panel d shows a Venn diagram identifying the number of differentially expressed genes identified by two or more methods versus uniquely by each.

939 **FIGURE 1**



940

FIGURE 2

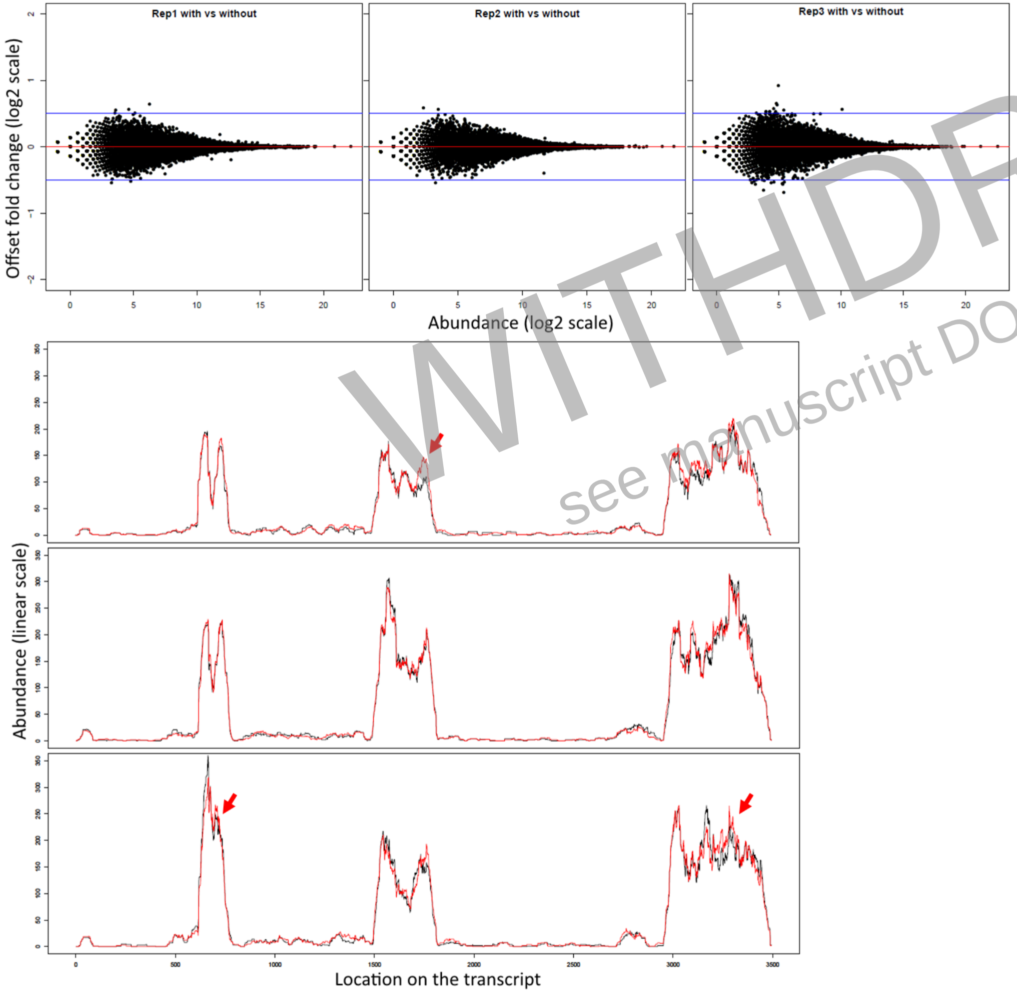

Rep1 with vs without
Rep2 with vs without
Rep3 with vs without
Offset fold change (log2 scale)
2
1
0
-1
-2
0
5
10
15
20
Abundance (log2 scale)
Abundance (linear scale)
0
50
100
150
200
250
300
350
0
500
1000
1500
2000
2500
3000
3500
Location on the transcript

FIGURE 3

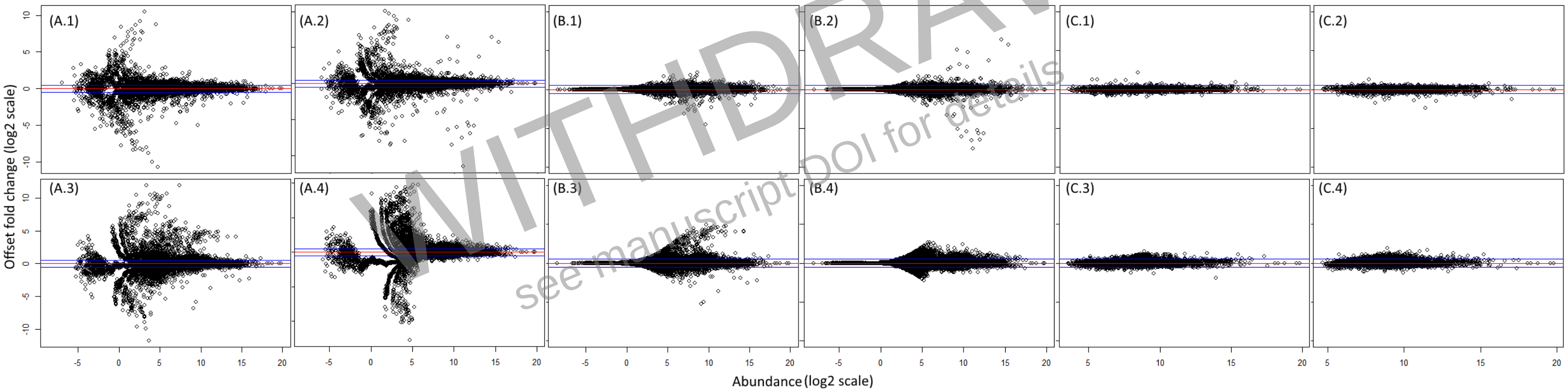
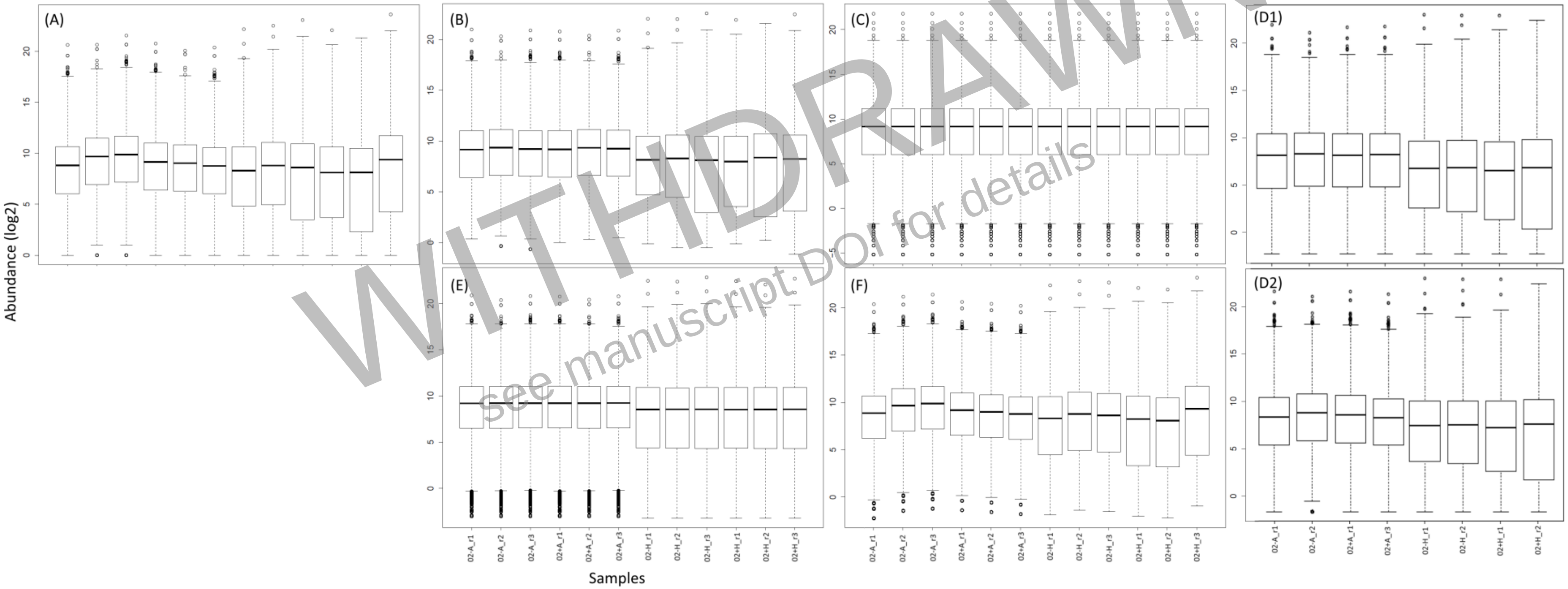
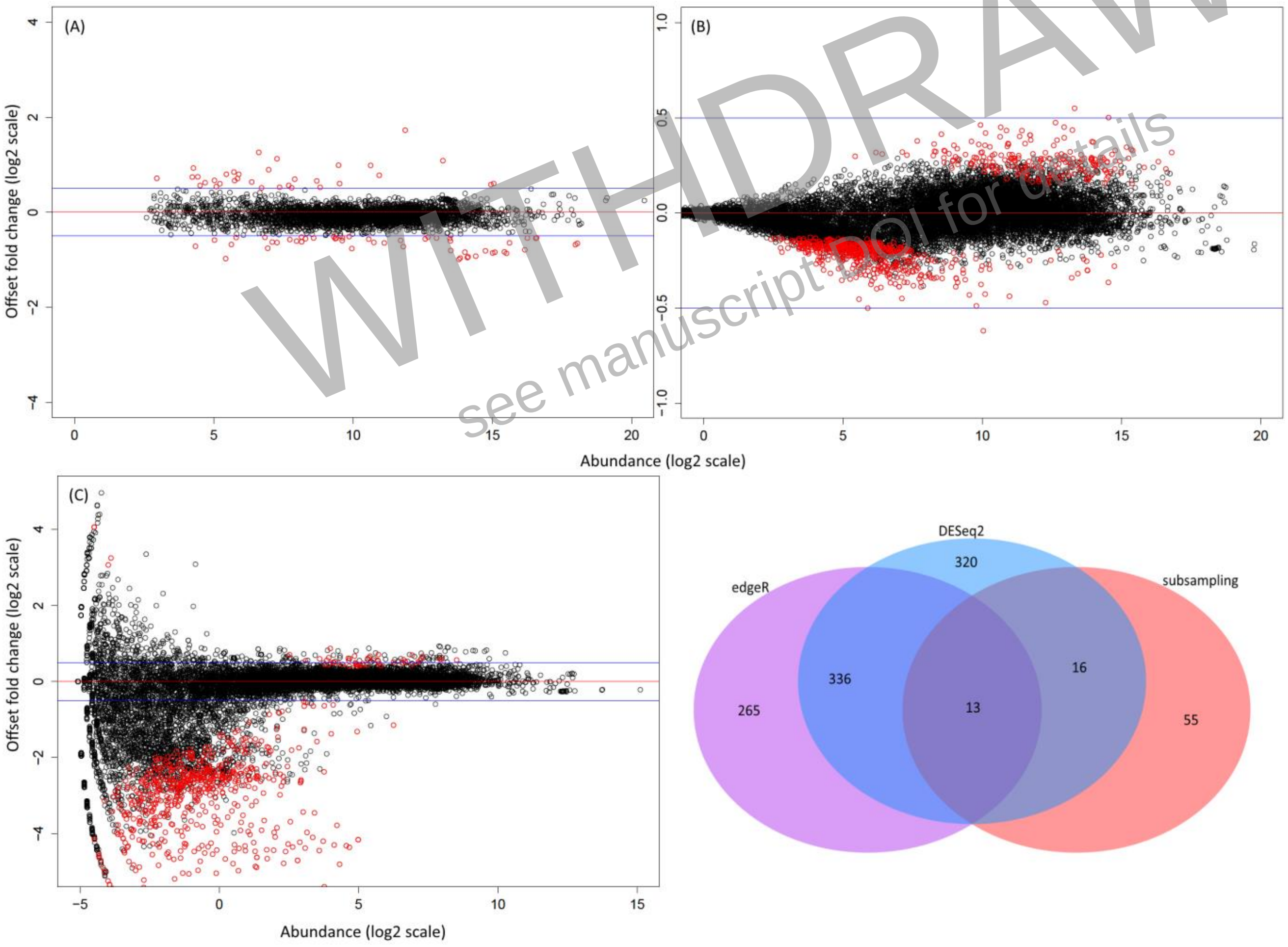

(A.1)
(A.2)
(B.1)
(B.2)
(C.1)
(C.2)
(A.3)
(A.4)
(B.3)
(B.4)
(C.3)
(C.4)
Offset fold change (log2 scale)
10
5
0
-5
-10
-5
0
5
10
15
20
Abundance (log2 scale)

FIGURE 4


(A)
(B)
(C)
(D1)
(E)
(F)
(D2)
Abundance (log2)
Samples
02-A_r1
02-A_r2
02-A_r3
02+A_r1
02+A_r2
02+A_r3
02-H_r1
02-H_r2
02-H_r3
02+H_r1
02+H_r2
02+H_r3

965 **FIGURE 5**


(A)
(B)
(C)
Offset fold change (log2 scale)
Abundance (log2 scale)
4
2
0
-2
-4
1.0
0.5
0.0
-0.5
-1.0
0
5
10
15
20
-5
DESeq2
edgeR
subsampling
320
336
16
13
265
55