A peer-reviewed version of this preprint was published in PeerJ on 29 October 2015.

Khang TF, Lau CY. 2015. Getting the most out of RNA-seq data analysis. PeerJ 3:e1360 https://doi.org/10.7717/peerj.1360

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, 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 response, 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 response of interest in the different phenotype classes 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 such as NOISeq or GFOLD. In contrast, for strong biological response, differentially expressed genes mined from unreplicated experiments using NOISeq, ASC and GFOLD had between 30 to 50% mean positive predictive value, an increase of more than 30-fold compared to the case of mild biological response. Among methods with good positive predictive value performance, having triplicates or more substantially improved mean positive predictive value to over 90% for GFOLD, 60% for DESeq2, 50% for NOISeq, and 30% for edgeR. We found DESeq2 to be the most reasonable method to call differentially expressed genes for systems level analysis as it showed the best PPV and sensitivity trade-off (mean PPV and mean sensitivity ~ 65% at replicate size of six).

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 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, 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 response, 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 response of interest in the different phenotype classes 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 such as NOISeq or GFOLD. In contrast, for strong biological response, differentially expressed genes mined from unreplicated experiments using NOISeq, ASC and GFOLD had between 30 to 50% mean positive predictive value, an increase of more than 30-fold compared to the case of mild biological response. Among methods with good positive predictive value performance, having triplicates or more substantially improved mean positive predictive value to over 90% for GFOLD, 60% for DESeq2, 50% for NOISeq, and 30% for edgeR. We found DESeq2 to be the most reasonable method to call differentially expressed genes for systems level analysis as it showed the best PPV and sensitivity trade-off (mean PPV and mean sensitivity $\sim$ 65% at replicate size of six).
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 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 thorugh 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, DE 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. Two recent studies compared the relative performance of large number of DEG call methods (eleven in Soneson and Delorenzi (2013); eight in Seyednasrollah et al. (2015)) under the R environment, and offerred recommendations for method selection. Nevertheless, it is important to keep in mind that the conclusions from the comparative studies were mostly derived from simulations based on synthetic data. More crucially, variation of the performance of DEG calling methods was not considered in the context biological effect and replicate size, which is 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. Nevertheless, some (e.g. NOISeq (Tarazona et al., 2011), GFOLD (Feng et al., 2012)) have options to deal with cases of unreplicated experiments. A minority such as ASC (Wu et al., 2010) is specifically designed for unreplicated 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

We investigated the performance of seven DEG call methods: GFOLD, ASC, NOISeq, edgeR (Robinson et al., 2010), DESeq2(Love et al., 2014), DESeq (Anders and Wolfgang, 2010) and Z-test. The first two are Bayesian methods, and were proposed to specifically address analysis of unreplicated RNA-seq data. However, GFOLD also has option to handle experiments with biological replicates. A popular nonparametric method is NOISeq, while edgeR and DESeq (and its updated version DESeq2) are commonly used parametric methods that explicitly model the distribution of count data using negative binomial distribution. Initially designed for standard experiments with biological replicates, these methods have been modified to accomodate analysis of unreplicated experiments, but their performance relative to

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1198v1 | CC-BY 4.0 Open Access | rec: 26 Jun 2015, publ: 26 Jun 2015

GFOLD and ASC remains unclear. The Z-test for equality of proportions was primarily used to set upper bounds in the tested performance metrics that are attainable by naive application of a common textbook statistical method. Specifically, the Z-test statistic for the ith gene is given by

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

where $\hat{p}_{ij}$ is the estimated proportion of the ith gene in the jth phenotype class ($j = 1,2$), $\hat{p}$ is the estimated pooled proportion of the ith gene, and N is the total number of normalized counts.

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 (Feng et al., 2012), 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 for fold change of. 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 requires the Linux platform, the other methods were implemented in R version 3.1.3 (R Core Team, 2015).

Benchmarking

To set up our benchmarking exercise, we needed two RNA-seq data sets whereby variation in their phenotype classes corresponded to mild and strong biological effect sizes in the tissue of interest. 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 $\sim$30 bases) obtained from the brain striatum tissues of two mice strains: C57BL/6J ($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 Wolfgang, 2010), which has been shown to be robust to library size and composition variation (Dillies et al., 2013). The exception is DESeq2, which specifically requires raw counts instead of normalized count for analysis.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1198v1 | CC-BY 4.0 Open Access | rec: 26 Jun 2015, publ: 26 Jun 2015

We proceeded to create an in silico gold standard set of DEG for each of the two RNA-seq data sets. To avoid biasing results of the called DEG due to algorithmic similarities, we decided to use the voom algorithm (Law et al., 2014; Ritchie et al., 2015). 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). In addition, a recent study reported that voom calls highly reliable DEG (Soneson and Delorenzi, 2013). 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 gold standard as it returned different DEG sets for different random seeds and number of permutation parameters (Supplemental Material Fig. S1).

Ideally, the in silico gold standard DEG called using voom should be 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 The remainder of the DEG that were unique to RNA-seq may either be false positives, or DEG that could not be detected using microarrays. Assuming that at most half of them were false positives, at least 70% of the voom-called DEG were expected to be real.

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 gold standard. 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 (based on BioGRID; Stark et al. (2006); Chatr-Aryamontri et al. (2015)) with at least one gene located on sex chromosomes (Supplemental Material Table T2).

Simulation and performance evaluation

To simulate unreplicated experiments in both data sets, we considered all possible paired samples from different phenotype classes. We discovered that the ASC package provided by (Wu et al., 2010) failed to run for particular combinations of sample pairs. As a result, only 27 and 124 pairs of samples from the Bottomly and Cheung data set respectively could be used for comparison across all methods. 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; the complement of the false discovery rate). For each DEG call method, we computed sensitivity as the proportion of gold standard DEG that were called. PPV was computed as the proportion of DEG called that were members of the set of gold standard DEG. The mean and standard deviation of these metrics were then reported. Methods that show good PPV are particularly interesting in the context of unreplicated experiments, since DEG candidates 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

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

Positive predictive value and sensitivity

The simulation results show that optimality of a DEG call method for a given replicate size depended on whether biological response was mild or strong (Fig.2). In the Cheung data set (mild biological response), all methods had very low (about 1%) mean positive predictive value (PPV) for unreplicated design (Supplemental Material Table T3). However, mean PPV ($\pm$SD) increased substantially for NOISeq

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1198v1 | CC-BY 4.0 Open Access | rec: 26 Jun 2015, publ: 26 Jun 2015

to $43.5 \pm 31.5\%$, and for GFOLD to $29.6 \pm 15.8\%$, for a design with $n = 3$. Doubling and approximately tripling the latter to $n = 6$ and $n = 10$ (Supplemental Material Fig. S2) 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 designs, mean PPV was low for the other methods; it did not exceed 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 response). In unreplicated experiments, mean PPV was relatively high for NOISeq ($47.9 \pm 23.7\%$), ASC ($47.2 \pm 25.9\%$) and GFOLD ($33.7 \pm 27.7\%$), compared to just about 15% in edgeR and 5% in DESeq and Z-test. Although DESeq2 showed reasonable mean PPV as well ($37.3 \pm 34.7\%$), the mean size of DEG called was very small ($6 \pm 11.6\%$). Interestingly, GFOLD attained very high mean PPV at $n = 3$ ($94.3 \pm 6.9\%$), 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. At $n = 3$ and $n = 6$, it 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. Moreover, at $n = 6$, DESeq2 had comparable 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 Liu et al.'s (Liu et al., 2014) study of the effect of replicate size (unreplicated experiments excluded) and sequencing depth, that statistical power primarily increases as a result of increasing biological replicate size.

DEG set size

Figure 3 shows the distribution of DEG set size in the Cheung and Bottomly data sets for different replicate sizes. Although DESeq2 could be used to call DEG for unreplicated experiments, it tended to make the least number of calls among methods, thus affecting its sensitivity. Because of this, its use in such cases does not seem justified. 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 response is strong, we suggest that DESeq2 is most suitable to mine DEG for systems biology work, on account of its good PPV and sensitivity balance. It should definitely replace DESeq.

Research programs focusing on investigation of weak or modest biological responses 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 both GFOLD and NOISeq at $n = 10$ became too small (below 20).

Table 1 summarizes the recommended DEG call methods and research goals for the combinations of biological effect size and replicate size considered in the present study.

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 rat, relative transcriptome coverage was about three times larger in the Cheung data set (human) compared to the Bottomly data set (rat). Despite this, detection of DEG remained difficult when biological effect size

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1198v1 | CC-BY 4.0 Open Access | rec: 26 Jun 2015, publ: 26 Jun 2015

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

Table 1. 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.

was weak, suggesting that the effect of transcriptome coverage on DEG calling was probably marginal in the present study.

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 biological response 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. Morever, systems level analysis remains difficult as none of the methods considered presently showed satisfactory sensitivity and positive predictive value performance. When strong biological response is expected, analysis of unreplicated experiments using GFOLD or NOISeq can yield DEG candidates with optimistic validation prospects. A standard triplicate design should result in further improvement. DESeq2 seems to be most suited for calling DEG for subsequent systems level analysis as it showed the best compromise between PPV and sensitivity among all tested methods.

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 (to be uploaded pending editorial decision).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1198v1 | CC-BY 4.0 Open Access | rec: 26 Jun 2015, publ: 26 Jun 2015

Funding

The study was supported by the University of Malaya Research Grant number UMRG RP032D-15AFR to T.F.K. The funders had no role in study design, data analysis, decision to publish, or preparation of the manuscript.

FIGURES

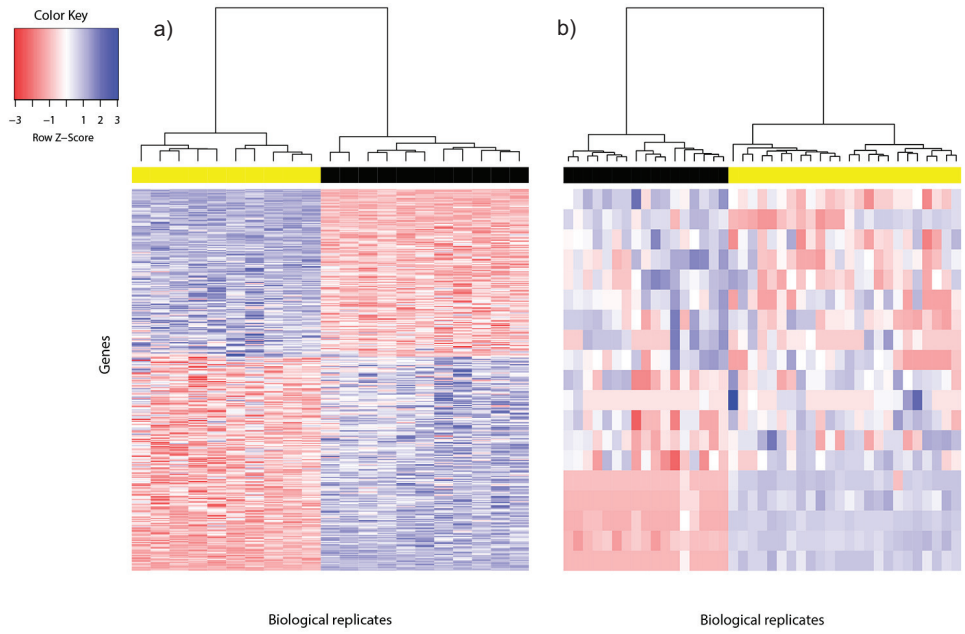

Figure 1. Heat map of differentially expressed genes in a) Bottomly data set (362 DEG) and b) Cheung data set (19 DEG). Phenotype class legend: a) Black for DBA/2J strain ($n = 11$); yellow for C57BL/6J strain ($n = 10$). b) Black for male ($n = 17$); yellow for female ($n = 24$). The heat maps were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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1198v1 | CC-BY 4.0 Open Access | rec: 26 Jun 2015, publ: 26 Jun 2015

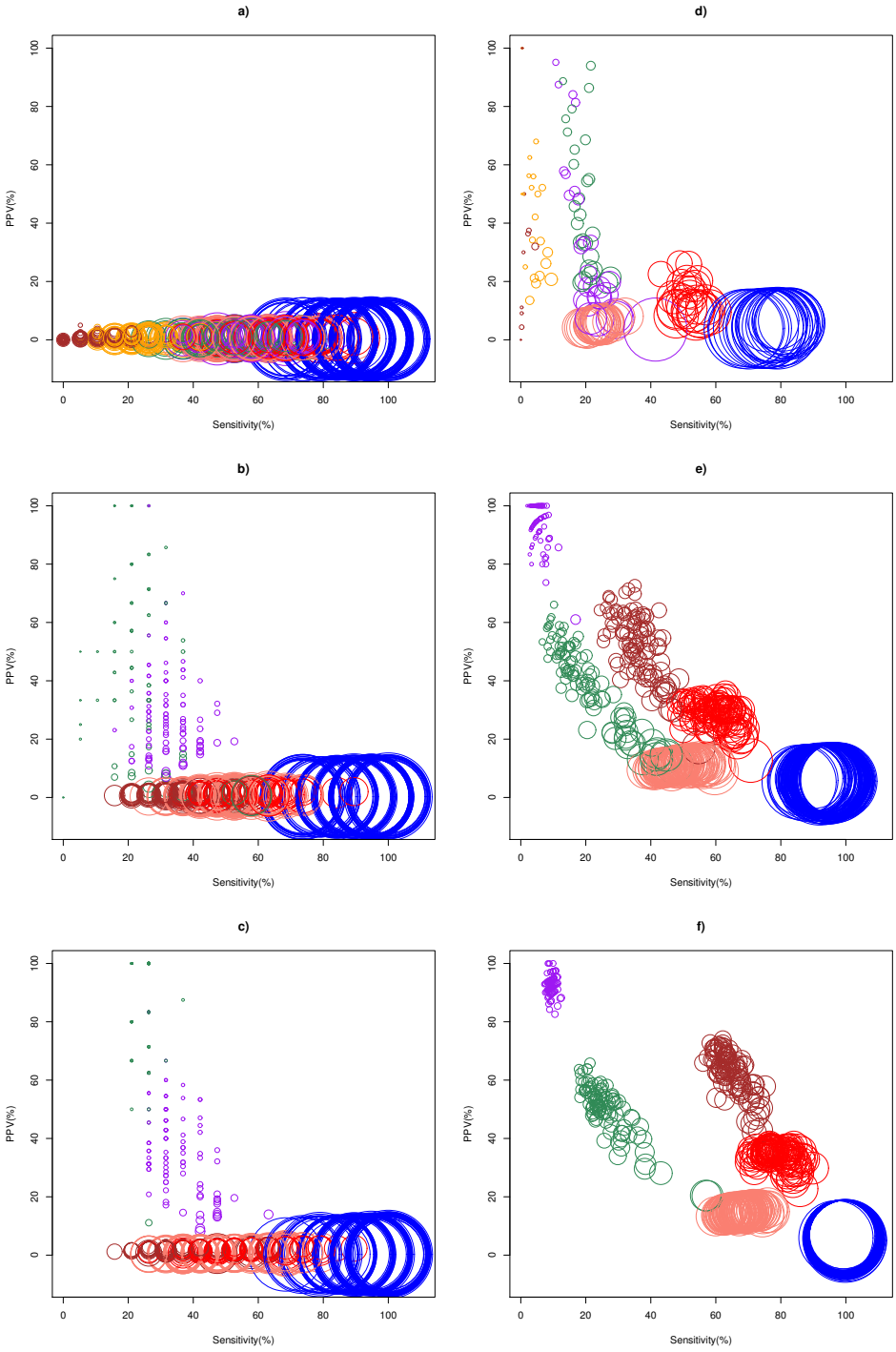

Figure 2. Scatter plots of PPV against sensitivity. The $n = 1, 3, 6$ scenarios are given in panels a,b,c for the Cheung data set, and d,e,f for the Bottomly data set, respectively. The diameter of a circle indicates the DEG set size. Color legend: blue(Z-test), pink(DESeq), red(edgeR), brown(DESeq2), purple(GFOLD), green(NOISeq), orange(ASC). For $n = 10$ in the Cheung data set, see Supplemental Material Fig. S2.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1198v1 | CC-BY 4.0 Open Access | rec: 26 Jun 2015, publ: 26 Jun 2015

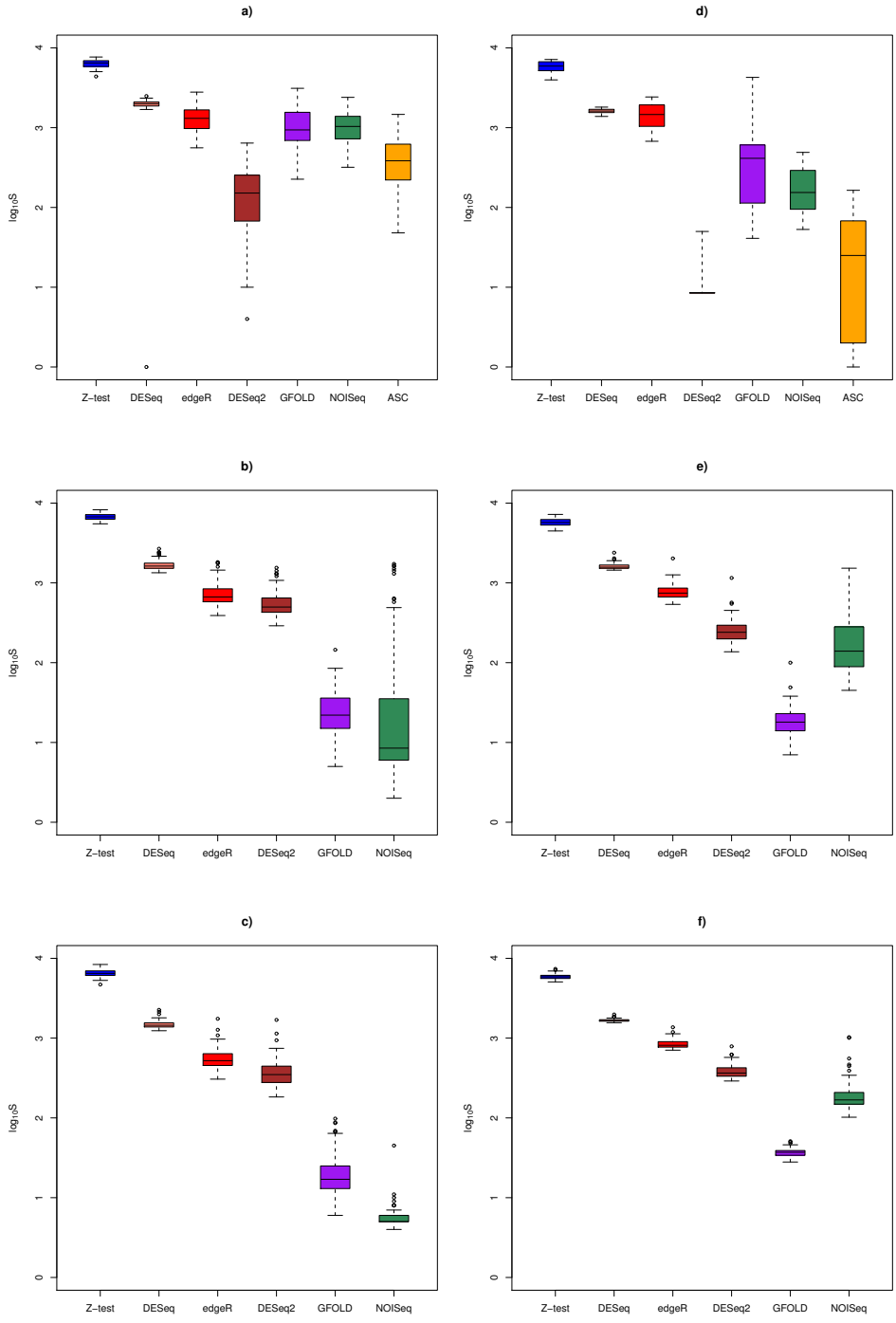

Figure 3. Box plots of distribution of DEG set size (in $\log_{10}$ scale) by method. The $n = 1, 3, 6$ scenarios are given in panels a,b,c for the Cheung data set, and d,e,f for the Bottomly data set, respectively. Color legend: blue(Z-test), pink(DESeq), red(edgeR), brown(DESeq2), purple(GFOLD), green(NOISeq), orange(ASC). For $n = 10$ in the Cheung data set, see Supplemental Material Fig. S3.

REFERENCES

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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1198v1 | CC-BY 4.0 Open Access | rec: 26 Jun 2015, publ: 26 Jun 2015

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

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.

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.

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.

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.

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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1198v1 | CC-BY 4.0 Open Access | rec: 26 Jun 2015, publ: 26 Jun 2015

differential expression in RNA-Seq data. *Statistical Methods in Medical Research*, 22:519–536.
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.
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., 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.
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, 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, 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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1198v1 | CC-BY 4.0 Open Access | rec: 26 Jun 2015, publ: 26 Jun 2015

significance score for gene selection and ranking. *Bioinformatics*, 30:801–807.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1198v1 | CC-BY 4.0 Open Access | rec: 26 Jun 2015, publ: 26 Jun 2015