Evaluation of read count based RNAseq analysis methods

Yan Guo*,†, Chung-I Li†, Fei Ye, Yu Shyr*

From The International Conference on Intelligent Biology and Medicine (ICIBM 2013)
Nashville, TN, USA. 11-13 August 2013

Abstract

Background: RNAseq technology is replacing microarray technology as the tool of choice for gene expression profiling. While providing much richer data than microarray, analysis of RNAseq data has been much more challenging. To date, there has not been a consensus on the best approach for conducting robust RNAseq analysis.

Results: In this study, we designed a thorough experiment to evaluate six read count-based RNAseq analysis methods (DESeq, DEGseq, edgeR, NBPSeq, TSPM and baySeq) using both real and simulated data. We found the six methods produce similar fold changes and reasonable overlapping of differentially expressed genes based on p-values. However, all six methods suffer from over-sensitivity.

Conclusions: Based on the evaluation of runtime using real data and area under the receiver operating characteristic curve (AUC-ROC) using simulated data, we found that edgeR achieves a better balance between speed and accuracy than the other methods.

Background

The process by which information from a gene is used in the synthesis of a functional gene product is called gene expression. The process of gene expression is used by all known life including eukaryotes, prokaryotes, and viruses to generate the macromolecular machinery for life. Gene expression analysis is essential for biomedical research.

The introduction of microarray technology has helped biomedical research make significant advances in the last decade by allowing high-throughput gene expression screening on all known genes. Recently, the introduction of RNAseq technology has had a revolutionary impact on the field of expression research. RNAseq refers to the use of next-generation sequencing (NGS) technologies to sequence cDNA in order to get information about a sample’s RNA content. Compared to microarray technology, the RNAseq method offers several distinct advantages. First, the detection range of RNAseq is not limited to a set of predetermined probes as with microarray technology, so RNAseq is capable of identifying new genes. Second, the resolution of a microarray is limited to the gene level for most arrays and the exon level for specially designed exon arrays. On the other hand, RNAseq can detect expression at the gene, exon, transcript, and coding DNA sequence (CDS) levels. Finally and most importantly, RNAseq can detect structural variants such as alternative splicing and gene fusion. With the maturity of NGS technologies, the price of RNAseq has become comparable to microarrays. Many researchers have predicted the inevitable replacement of microarray by RNAseq [1-3] based on the competitive price and additional genomic information contained in RNAseq data.

With the rich genomic information RNAseq technology brings, it also comes with complication in the analysis phase. To date, the research community has not yet come to a consensus on the best approach for analyzing RNAseq data. The most popular normalization method for microarray data is Robust Multi-array Average (RMA) [4], a form of quantile normalization. For RNAseq, one of the popular normalization methods is Reads per Kilobase per Million mapped reads (RPKM) [5] or Fragments Per Kilobase of transcript per Million mapped reads (FPKM) [6]. The RPKM of a gene

* Correspondence: yan.guo@vanderbilt.edu; yu.shyr@Vanderbilt.Edu
† Contributed equally
Center for Quantitative Sciences, Vanderbilt University, Nashville, TN, USA

© 2013 Guo et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
is computed as follows: $\text{RPKM} = 10^9 \times \frac{C}{N \times L}$, where $C$ is the number of reads mapped to the gene, $N$ is the total number of reads mapped to all genes, and $L$ is the length of the gene. FPKM is computed similarly to RPKM, except it accounts for the scenario in which only 1 end of a paired-end read is mapped. In addition to RPKM and FPKM, other read count methods based on Poisson, negative binomial, and Bayesian methods also exist. Each method has unique strengths and weaknesses. In this study, we focus on read count-based methods and systematically evaluate 6 RNAseq R packages including DESeq [7], DEGseq [8], edgeR [9], baySeq [10], TSPM [11] and NBPSeq [12] using both real and simulated data. BaySeq is considered an empirical Bayes approach to detect patterns of differential expression, DESeq and NBPSeq are based on a negative binomial model, DEGseq and TSPM are based on a Poisson model, and edgeR uses empirical Bayes estimation and exact tests based on the negative binomial.

**Method**

The real RNAseq datasets were selected from The Cancer Genome Atlas (TCGA) [13,14]. TCGA is a massive, comprehensive, and collaborative project to catalogue genomic data for over 20 types of cancers by the National Cancer Institute (NCI), the National Human Genome Research Institute (NHGRI), and 27 institutes and centers of the National Institute of Health (NIH). Gene expression profiling by RNAseq is one of the major components of genomic data collected by TCGA. Breast cancer is the only cancer type in TCGA that collected expression data on a large quantity of tumor-normal paired samples. Thus we selected breast cancer tumor-normal paired data (53 pairs) as our primary source of real RNAseq data. Differentially expressed genes between tumor and normal were identified using all six methods at the significance level of 0.05 with Benjamin-Hochberg False discovery rate (BH FDR) adjustment. To evaluate the consistency between the six methods, we computed pairwise Spearman’s correlations as well as intraclass correlation (ICC) for fold change values of all genes, along with the corresponding p-values. In addition, we evaluated each method’s running time.

For each gene, the count is drawn from the negative binomial distribution with the mean and dispersion parameters estimated from the TCGA breast cancer dataset. For a given gene $m$, the fold change is calculated as $(\rho_m + \rho_c)^{dm}$, where $\rho_m$ is drawn from the gamma distribution with shape parameter 0.87 and rate parameter 1.36 (parameters are estimated from the TCGA breast cancer dataset), $\rho_c$ is the lower bound of fold change, and

$$dm = \begin{cases} 1, & \text{if gene } m \text{ is upregulated} \\ -1, & \text{if gene } m \text{ is downregulated} \\ 0, & \text{otherwise} \end{cases}$$

We evaluated the methods using datasets simulated to present different scenarios corresponding to a given combination of the following parameters: sample size (5 or 10), proportion of differentially expressed genes (5% or 10%), ratio of up-regulated vs. down-regulated (1:1 or 3:1), lower bound of fold change (1.5 or 1.1), and lower bound of depth (5 or 1).

In order to measure the overall performance, we also computed the area under the curve (AUC) across 30 simulation datasets under each scenario. False discovery rate (FDR) at levels of 0.1, 0.05 and 0.01 were used as the cutoffs. In order to measure the overall performance, we also computed the area under the curve (AUC) across 30 simulation datasets under each scenario.

**Results and discussion**

We performed differential expression analysis on the 53 tumor-normal paired samples from TCGA using all six methods. Out of 16,146 genes, DEGseq identified the most number of significant genes at 15,226. That is, 94.3% of all genes were identified by DEGseq as statistically significantly differentially expressed between tumor and normal, which implies that DEGseq is over-sensitive. DESeq on the other hand identified the least number of differentially expressed genes with 7,171, which is still too many to be biologically plausible. NBPSeq, edgeR, TSPM, and baySeq identified 10,017, 10,457, 9,519, and 13,203 differentially expressed genes respectively (Figure 1). In terms of up-regulated genes vs. down-regulated genes, DESeq, edgeR, and NBPSeq identified more up-regulated genes, while TSPM and DEGseq identified more down-regulated genes. BaySeq does not provide fold change or test statistics, thus no direction of gene expression change can be determined through p-value alone. A unique characteristic associated with baySeq is that there is a randomization factor built into its computation model. Using baySeq to analyze the same dataset on same parameter settings twice will generate slightly different results. In our scenario, the difference can be as many as 100 genes.

Next we computed the overlap of differentially expressed genes between methods for both up-regulated and down-regulated genes. BaySeq was excluded from this analysis due to its lack of directionality (Figure 2). For down-regulated genes, 2,521 were identified by the rest of five methods, and for up-regulated genes, 4,067 were identified by all five methods. DESeq, edgeR, and NBPSeq observed the fewest singleton genes (defined as genes identified by only 1 method) in terms of both up-regulated and down-regulated genes. TSPM observed a moderate number of singleton genes, and DEGseq
observed the most, which is also a reflection of its oversensitivity.

We studied the relationship of fold change and p-value vs. identification frequency of significantly expressed genes detected by the six methods (Figure 3). As expected, genes identified by all six methods had a stronger signal (e.g., a higher fold change value) than genes identified by fewer methods. The singleton genes had the lowest fold change values (Figure 3 left). However, for p-value the pattern was not clear. No significant association between p-value and identification frequency was observed.

We also tested the consistency among the six methods at overall significance levels. Pair-wise Spearman’s correlation coefficient of fold change values was computed (Figure 4). Most correlation coefficients are very close to 1, which indicates that the six methods are highly consistent in terms of ranking genes according to fold change despite difference in normalization method. We also computed intraclass correlation coefficients (ICC) using both fold change and p-value as well. The ICC based on fold change was 0.975 which strongly suggests that a high degree of agreement among the five methods (excluding baySeq because no fold change can be calculated) as the fold change value is only affected by different normalization techniques used by different methods (Figure 5) However, the ICC for p-value was only 0.50, indicating that these methods have a low agreement in terms of ranking the differentially expressed genes. In addition, baySeq produced the worst p-value correlation with regard to the other methods. To assess accuracy, one has to know a priori which genes are differentially expressed between normal and tumor. Hence, we can only calculate it in a simulation.

For simulated data, we compared the number of genes that are significantly differentially expressed. In contrast to the results from real data, baySeq identified the smallest number of differentially expressed genes, while DEGseq identified largest number of differentially expressed genes, significantly more than other methods (Table 2). DESeq, edgeR and NBSeq identified similar numbers of differentially expressed genes. The similar performance of DESeq, edgeR, and NBSeq is probably due to the fact that these methods are based on similar principles except the estimation of the dispersion parameter. Among all six methods, baySeq has the smallest FPR, and DEGseq has the largest FPR. This indicates that baySeq was more conservative compared to other methods. Among DESeq, edgeR and NBSeq, the largest FPR was found by NBSeq, followed by DESeq or edgeR. In general, the FPR of DESeq was smaller than the FPR of edgeR. For TPR, the largest TPR was found by DEGseq (i.e. it also found too many false positives), followed by TSPM. As expected, DESeq, edgeR and NBSeq obtained similar results in TPR. After comparing the results under scenarios I and II, we found that sample size has a large effect, especially for TPR. As sample size decreases, N decreases, FPR increases, and TPR decreases, respectively. TSPM shows the strongest effects on sample size among the methods. After comparing the results under scenarios I and III, we found that the proportion of differentially expressed genes has a relatively small effect on FPR and TPR for all methods. As the proportion decreases, all of N decreases, FPR, and TPR decrease. After comparing the results under scenarios I and IV, we found that the proportions of up-regulated and down-regulated genes have a large effect on N and FPR for TSPM and NBSeq. After comparing the results under scenarios I, V, and VI, we found that the level of treatment effects and depth of reads have small effects for all of the methods under reasonable sample size. Comparing the results

| Table 1 Scenarios in the simulation comparison study |
|-----------------|--------|-----|------|-----------------|--------|
| Scenario | Sample size | DE (%) | Up/down | Lower bound of fold change | Lower bound of depth |
| I | 10 | 10 | 1:1 | 1.5 | 5 |
| II | 2 | 10 | 1:1 | 1.5 | 5 |
| III | 10 | 5 | 1:1 | 1.5 | 5 |
| IV | 10 | 10 | 3:1 | 1.5 | 5 |
| V | 10 | 10 | 1:1 | 1.1 | 5 |
| VI | 10 | 10 | 1:1 | 1.5 | 1 |
| VII | 2 | 10 | 1:1 | 1.1 | 1 |

Figure 1 Number of significantly differentially expressed up-regulated and down-regulated genes for each method.
under scenarios I and VII, all methods are sensitive to the sample size, level of treatment effects, and depth.

In Table 3, we summarized the average AUC-ROC for each scenario. For each scenario, DEGseq has less satisfying performance compared to the other methods. Under scenarios II, V, and VII, the average AUC-ROC of baySeq was the largest. Those three scenarios were simulated for the dataset with small sample size or small treatment effects. This indicates that baySeq is the best performing method when the dataset has a small sample size or small treatment effect. Under other scenarios (i.e. I, III, IV, VI), the best performance was generally obtained by edgeR and DESeq, followed by NBPSeq and TSPM.

We also measured the runtime for each method running on the TCGA breast cancer dataset. On a Dell 3500 with a 2.8GHz CPU and 12 GB RAM, DESeq took the most time to finish with over 2 hours followed by NBPSeq (77 minutes), baySeq (47 minutes). TSPM and edgeR took significantly less time in comparison with 6 and 2 minutes, respectively. DEGseq was the speediest of the six, finishing in only 12 seconds. Combining information from both runtime and AUC, edgeR performed best with relatively short runtime and good AUC-ROC.

**Conclusions**

In this study, we systematically evaluated six read count-based RNAseq analysis methods using both real and simulated data. BaySeq is the most unique method out of the six, because it is based on an empirical Bayesian approach which produces no fold change or test statistics.
to infer the direction of the gene expression difference. This inconvenient feature of baySeq forces users to go back to the raw expression value to determine the direction. Another unique characteristic of baySeq is the randomization steps involved in its model. For the same data with the same parameter settings, baySeq produces slightly different results. Granted, genes with the strongest fold change will always be detected by baySeq, but this minor inconsistency between runs of baySeq does produce some inconvenience. From our simulation study, we found that the performances of DESeq, edgeR, and NBPSeq are similar in most cases. This result is expected because all three methods are based on a negative binomial model and differ principally in the way the dispersion parameter is estimated.

One of the issues we observed with these six read count-based RNAseq analysis methods is that they tend to be over-sensitive. For a traditional microarray study, the number of differentially expressed genes identified by a simple method such as a t-test is highly dependent on the sample size. We have tested these six methods using smaller sample sizes such as 2 vs. 2, but the majority of the methods still produced huge amounts of differentially expressed genes (sometimes over 50% of total genes), which is a clear sign of over-sensitivity. Through more thorough analysis, we found that the five methods (excluding baySeq) produced very similar fold changes but less similar ranks of p-value. Given these facts, we would recommend using the combined criteria of fold change and p-value to filter out more false positives. Combined with evaluation of runtime from real data and AUC-ROC from simulated data, edgeR achieved a good balance between speed and accuracy.

We are far from reaching a consensus on the best RNAseq analysis approach. Several unique characteristics of RNAseq data contribute to the difficulty of RNAseq data analysis. The value for non-expression in RNAseq is zero, which means there are no reads aligned to the gene. In microarray, there is always background intensity for non-expressed genes, thus we can always take log transformations of the intensity data. In contrast, due to the large number of zeros in RNAseq data (often more than 50%), we cannot take a log transformation. The range of RNAseq data is between 0 and 50,000, compared to microarray’s 2 to 15 after RMA normalization. This huge range of RNAseq data can result in many false high fold changes. Also, there are many sequencing and alignment artifacts that can skew
### Figure 5
Intraclass correlation coefficients among six methods using fold change and p-value

### Table 2 Performance comparison under different scenarios

|     | FDR = 0.1 | FDR = 0.05 | FDR = 0.01 |
|-----|-----------|------------|-----------|
|     | Ns, FPR, TPR | Ns, FPR, TPR | Ns, FPR, TPR |
| I   | DESeq     | 759, 0.0090, 0.677 | 679, 0.0054, 0.630 | 550, 0.0020, 0.531 |
|     | Edger     | 825, 0.0115, 0.721 | 740, 0.0068, 0.679 | 614, 0.0023, 0.593 |
|     | DESeq     | 8814, 0.873, 0.953 | 8864, 0.857, 0.857 | 8415, 0.830, 0.940 |
|     | NBSeq     | 1167, 0.0480, 0.734 | 972, 0.0322, 0.682 | 723, 0.0153, 0.585 |
|     | BaySeq    | 729, 0.0060, 0.675 | 654, 0.0027, 0.629 | 543, 0.0006, 0.538 |
|     | TSPM      | 2307, 0.168, 0.789 | 1815, 0.0120, 0.730 | 1202, 0.0663, 0.605 |
| II  | DESeq     | 289, 0.0153, 0.152 | 175, 0.0096, 0.089 | 59, 0.0037, 0.025 |
|     | Edger     | 239, 0.0138, 0.114 | 130, 0.0080, 0.058 | 38, 0.0028, 0.012 |
|     | DESeq     | 8379, 0.829, 0.916 | 8215, 0.811, 0.908 | 7921, 0.780, 0.892 |
|     | NBSeq     | 308, 0.0207, 0.121 | 194, 0.0134, 0.072 | 70, 0.0054, 0.021 |
|     | BaySeq    | 189, 0.0070, 0.126 | 88, 0.0029, 0.061 | 10, 0.0003, 0.006 |
|     | TSPM      | 291, 0.0262, 0.055 | 262, 0.0234, 0.051 | 217, 0.0191, 0.044 |
| III | DESeq     | 363, 0.0056, 0.620 | 319, 0.0034, 0.573 | 255, 0.0013, 0.486 |
|     | Edger     | 401, 0.0068, 0.672 | 354, 0.0041, 0.630 | 284, 0.0014, 0.541 |
|     | DESeq     | 8451, 0.839, 0.959 | 8288, 0.822, 0.955 | 8002, 0.792, 0.946 |
|     | NBSeq     | 567, 0.0241, 0.677 | 474, 0.0169, 0.628 | 346, 0.0084, 0.533 |
|     | BaySeq    | 332, 0.0031, 0.605 | 295, 0.0015, 0.561 | 242, 0.0004, 0.476 |
|     | TSPM      | 851, 0.0529, 0.689 | 651, 0.0354, 0.631 | 426, 0.018, 0.503 |
| IV  | DESeq     | 758, 0.0090, 0.677 | 678, 0.0055, 0.629 | 544, 0.0019, 0.526 |
RNAseq data. For example, GC content plays an important factor in the sequenceability of a gene, and the exon’s length has a large effect on alignment accuracy. To date, only 1 RNAseq package, DESeq takes the paired data information into consideration. In summary, even though there are a large number of RNAseq analysis tools at our disposal, given the uniqueness of RNA-seq data and the number of unsolved problems, there is still much room left to improve RNAseq analysis.

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
Yan Guo and Chung-I performed the analysis and wrote the manuscript. Fei Ye and Yu Shyr provided significant scientific input. All authors read and approved the final manuscript.

Declarations
The publication fee of this manuscript is provided by Harold L. Moses Chair in Cancer Research fund, Vanderbilt University. This article has been published as part of BMC Genomics Volume 14 Supplement 8, 2013: Selected articles from the International Conference on Intelligent Biology and Medicine (ICIBM 2013): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/14/S8.

Published: 9 December 2013

References
1. Wang Z, Gerstein M, Snyder M: RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009, 10(1):57-63.
2. Shendure J: The beginning of the end for microarrays? Nat Methods 2008, 5(7):585-587.
3. Guo Y, Sheng Q, Li J, Ye F, Samuels DC, Shyr Y: Large Scale Comparison of Gene Expression Levels by Microarrays and RNASeq Using TCGA Data. PLoS One 2013, 8(8):e71462.
4. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: Exploration, normalization, and summaries of
5. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008, 5(7):621-628.

6. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 2010, 28(5):511-515.

7. Anders S, Huber W: Differential expression analysis for sequence count data. Genome Biol 2010, 11(10):R106.

8. Wang L, Feng Z, Wang X, Zhang X: DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 2010, 26(1):136-138.

9. Robinson MD, McCarthy DJ, Smyth GK: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010, 26(1):139-140.

10. Hardcastle TJ, Kelly KA: baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 2010, 11:422.

11. Auer L, Paul, RWD: A Two-Stage Poisson Model for Testing RNA-Seq Data. Statistical Applications in Genetics and Molecular Biology 2011, 10(1):26.

12. Di Y, SD, Cumbie JS, Chang JH: The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. Statistical Applications in Genetics and Molecular Biology 2011, 10(1).

13. Integrated genomic analyses of ovarian carcinoma. Nature 2011, 474(7353):609-615.

14. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012, 487(7407):330-337.

doi:10.1186/1471-2164-14-S8-S2

Cite this article as: Guo et al.: Evaluation of read count based RNA-seq analysis methods. BMC Genomics 2013 14(Suppl 8):S2.