Journal of Bioinformatics and Computational Biology
© Imperial College Press

CONDITIONAL ESTIMATION OF LOCAL POOLED DISPERSION PARAMETER IN SMA LL-SAMPLE RNA-SEQ DATA IMPROVES DIFFERENTIAL EXPRESSION TEST

JUNGSOO GIM

Institute of Health and Environment, Seoul National University, Gwanak-gu
Seoul, 151-747, South Korea
jgim80@snu.ac.kr

SUNGHO WON

Graduate School of Public Health, Seoul National University, Gwanak-gu
Seoul, 151-747, South Korea
won1@snu.ac.kr

TAESUNG PARK

Department of Statistics, Seoul National University, Gwanak-gu
Seoul, 151-747, South Korea
Corresponding author
tspark@stats.snu.ac.kr

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

High throughput sequencing technology in transcriptomics studies contribute to the understanding of gene regulation mechanism and its cellular function, but also increases a need for accurate statistical methods to assess quantitative differences between experiments. Many methods have been developed to account for the specifics of count data: non-normality, a dependence of the variance on the mean, and small sample size. Among them, the small number of samples in typical experiments is still a challenge. Here we present a method for differential analysis of count data, using conditional estimation of local-pooled dispersion parameters. A comprehensive evaluation of our proposed method in aspect of differential gene expression analysis using both simulated and real data sets describes that the proposed method is more powerful than other existing methods while controlling the false discovery rates. By introducing conditional estimation of local pooled dispersion parameters, we successfully overcome the limitation of small power and enable a powerful quantitative analysis focused on differential expression test with the small number of samples.

Keywords: Differential expression test; RNA-Seq analysis; Local pooled dispersion estimation; Small sample data

1. Background

There is considerable interest in biomedical research with high dynamic range and low background expression level accuracy in sequencing technology, such as RNA-Sequencing (RNA-Seq). To take advantage of RNA-Seq, a statistical method that is accurate and robust to a small number of sample size is required [1]. Specifically, knowing the differential transcription regulation pattern lies at the heart of many aspects of biomedical research and it is therefore desirable to understand its differential regulation patterns from a small number of samples, in typical RNA-Seq experiments as much as possible.

Many attempts, to improve our understanding of transcriptional difference between groups of interests, have been made by pooling information across genes while assuming similarity of dispersion (or variance) of different genes. Since Robinson and Smyth first introduced the negative binomial (NB) distribution to model a count structure of SAGE data [2], numerous NB-based models have been developed for RNA-Seq. *edgeR* assumed one common dispersion parameter that is common throughout the genes [3], while *DESeq* incorporated two parameters under the assumption of different per-gene dispersion [4]. Using shrinkage estimation of dispersions and log fold changes, Love *et. al.* developed *DESeq2* (as an extension of *DESeq*) to improve stability and interpretability of the estimates [5]. There also have been other approaches assuming different distributions, such as Poisson [6] and log-normal [7]. In addition to these parametric approaches, Bayesian approaches [8, 9] and non-parametric approaches [6, 10] were proposed to statistically detect the changes in expression between treatment and control groups.

Many methods adopt an approach of pooling across genes to overcome the limitation of small sample data analysis. Specifically, by exploiting the assumption of similarity of the dispersions of different genes measured in the same experiment, the dispersion parameter is estimated from such information shared across genes. A number of review papers reported, however, that a majority of these methods are limited by their poor performance when the sample size is small [1, 11, 12]. Thus there still remains a need for a powerful method that takes into account the properties of small samples with more caution.

In this paper we apply the idea of using local pooled error [13] and conditional likelihood method [2] together to develop a new statistical analysis method, called conditional estimation of local pooled dispersion parameter for RNA-Seq data or *cLPD-seq* which performs with stability when using small samples. Our method shares information over genes in each pre-defined local bin and makes a smoothing curve in order to stabilize dispersion estimation in small samples. We demonstrate the advantages of our method by analyzing simulated datasets and apply it to real RNA-Seq data

2. Materials and Methods

2.1. *Preliminaries*

Let y_{ij} denote the observed count for class i and library j for a single gene. Here we assume a two-class comparison so that $i = 1, 2$ and for a given i, $j = 1, ..., n_i$ which is required to be a natural number greater than or equal to 1. Since the results from comprehensive reviews indicate that the negative binomial (NB) modeling of RNA-Seq data performs relatively better than others, we model the gene counts y_{ij} as NB distributed,

$$Y_{ij} \sim NB(\mu_{ij} = \rho_i m_{ij}, \phi) \tag{1}$$

where ϕ is the dispersion parameter and $E[y_{ij}] = \mu_{ij}$ and $Var[y_{ij}] = \mu_{ij} + \mu_{ij}^2 \phi$. Let ρ_i denote the library size (or sequencing depth) and let m_{ij} be the true abundance of a gene of interest and the library size of sample j in class i , respectively.

2.2. *Common dispersion estimation*

We incorporate a common dispersion model developed by Robinson and Smyth [2], which uses all genes to estimate a common dispersion (ϕ) of SAGE data. The conditional likelihood for a gene is formed by conditioning on the sum of counts for each class. If the library size m_{ij} is equal within each class, the conditional log-likelihood for ϕ of a gene g, given $Y_i = \sum_{j=1}^{n_i} Y_{ij}$ is:

$$l_g(\phi) = \sum_{i=1}^{2} \left[\sum_{j=1}^{n_i} \log\left(\Gamma\left(y_{ij} + \phi^{-1}\right)\right) + \log\left(\Gamma\left(n_i \phi^{-1}\right)\right) - \log\left(\Gamma\left(Y_i + n_i \phi^{-1}\right)\right) - n_i \log\left(\Gamma\left(\phi^{-1}\right)\right) \right] \tag{2}$$

where $\Gamma(n)$ denotes the Gamma function, which is related to the factorial by $\Gamma(n) = (n-1)!$. The common dispersion estimator maximizes the common likelihood $l_C(\phi) = \sum_{g=1}^{G} l_g(\phi)$ where G is the number of genes. A quantile adjustment can be used to adjust for unequal library sizes, as is done in [2, 14].

2.3. *Gene-wise dispersion estimation*

The common dispersion assumption offers significant stabilization, compared to gene-wise estimation, especially with small samples [14]. However, this assumption that each gene has the same dispersion is not likely. Therefore, we suggest a 'local common dispersion' assumption, similarly suggested by others in microarray data analysis [13, 15]. We state our gene-wise dispersion parameter estimation strategy through local-bin estimation as follows.

(1) Evaluate average value of each gene across the samples, $\mathbf{A} = (A_1, ..., A_g, ..., A_G)$. Here $\mathbf{A}$ and A_g respectively denote the mean vector and the mean of gene g .

(2) Define q number of local bins using two different binning approaches: q-equal-frequency and q-equal-space (q = 100 and binning = equal-space as default). In equal-frequency mode, quantiles of $\mathbf{A}$ are evaluated and used to define local bins, while an equal-spaced distance, (max($\mathbf{A}$)-min($\mathbf{A}$))/q, is used to define local bins in equal-space approach.

(3) Place the genes into the bins where their average values belong to.

(4) For each bin k (where $k=1,\ldots,q$), find the common dispersion estimator $\hat{\phi}_k$ which maximizes l_C .

(5) Generating a smoothing curve (using local regression, or LOESS) with q number of $\hat{\phi}_k$ estimates. Note that other smoothing techniques can be applied.

(6) Gene-wise dispersion is estimated via interpolation method on this curve.

Note that two different binning approaches in step 2 might affect estimation of dispersion. Hence we computed the smoothing curves of both cases and compared the performances (see Fig. 1 for typical estimated curves).

2.4. *Statistical testing*

For DE test of small sample RNA-Seq data, we adopted the exact test used in *edgeR*, developed by Robinson and Smith [14]. For two-class comparisons, the count values of genes under the null of no difference are identically distributed, leading to known distributions of the within-condition count totals for each gene. Also, the sum of the total gene counts over all libraries has a known distribution. Let Y_1 and Y_2 be the sum of counts for class 1 and 2, respectively, over the number of libraries, n_1 and n_2 . An exact test similar to the Fisher's exact test for contingency tables can be constructed by replacing the hypergeometric probabilities with NB. Conditioning on the total count-sum, Y_1+Y_2 , which is also an NB variable, the probability of observing class totals at least as extreme as the

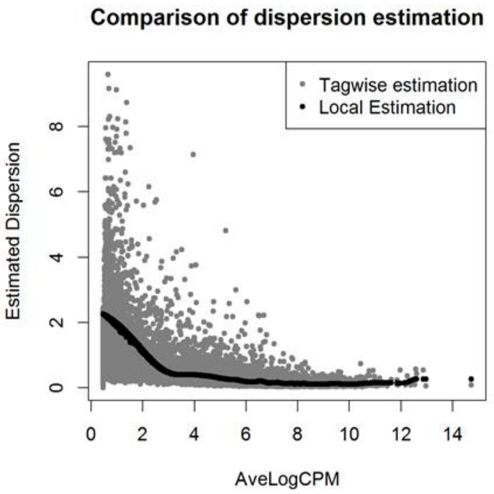

Fig. 1. Comparison of dispersion parameter estimation. Tag-wise estimates (based on weighted likelihood method in *edgeR*) and Local bin estimates (proposed method) are depicted in grey and black colour, respectively. X-axis and Y-axis represent average log-transformed counts per million (CPM) and estimated dispersion parameter values.

observed can be calculated, resulting in the exact p-value for differential expression. We used the R function implemented in *edgeR*.

2.5. *Data sets*

For a practical comparison of the methods, we analyzed both simulated data from Love *et al.* [5] and publicly available real data which had been preprocessed and distributed by Recount [16]. Brief characteristics of each dataset is introduced below.

- Simulated dataset (6, 8, 10, and 20 samples): Love *et al.* [5] used real data-based simulated dataset to benchmark a number of methods. Based on the reproducible code in [5], we generated simulated dataset. Briefly, 10 000 genes were randomly selected and samples were generated from NB distribution. Among those genes, 80% were generated from the null hypothesis, while the remaining 20% were generated from the alternative hypotheses having different fold changes (FC) between classes: 2 and 4 FC. The directions of FC were randomly assigned. The mean and dispersion values for this simulation were drawn from Pickrell *et al*. [17].
- Hapmap dataset (60 samples in normal condition): Montgomery *et al.* [18] performed sequencing the mRNA fraction of the transcriptome of hymphoblastoid cell lines (LCLs) from 60 CEU (HapMap individuals of European descent) individuals to understand the quantitative difference in gene expression within a human population.
- Gilad group dataset (6 samples): Gilad group [19] performed comparative studies to assess intra- and interspecies variation in gene regulatory processes. They used RNA-seq to study transcript levels in humans, chimpanzees, and rhesus macaques, using liver RNA samples from three males and three females from each species.
- Fly dataset (10 samples): Graveley *et al.* [20] studied 30 distinct developmental stages in *Drosophila melanogaster* using RNA-Seq, tiling microarrays and cDNA sequencing. Here we selected a part of the RNA-Seq experimental results with 10 samples in two different developmental stages.

2.6. *Method comparison*

We compared our proposed method with *edgeR* [3, 21], *DESeq* [4], and *DESeq2* [5] based on the review report of best performance of NB-based method in small sample studies [1, 11]. We also included non NB-based methods: *voom* [7] which assumes normality of log-transformed counts, and *SAMseq* [6] which is a non-parametric method. The Benjamini–Hochberg procedure was used to adjust multiple testing problem [22]. Transcripts were reported as DE at an adjusted p-value threshold of 0.1. We compared true positive rate (TPR, or sensitivity) and false positive rate (FPR, or 1-specificity) defined as the number of true DE transcripts detected divided by the number of true DE transcripts and the number of false DE transcripts detected divided by the number of non-DE transcripts, respectively. In most cases, we ran the programs using the settings provided in the supplementary material of [1].

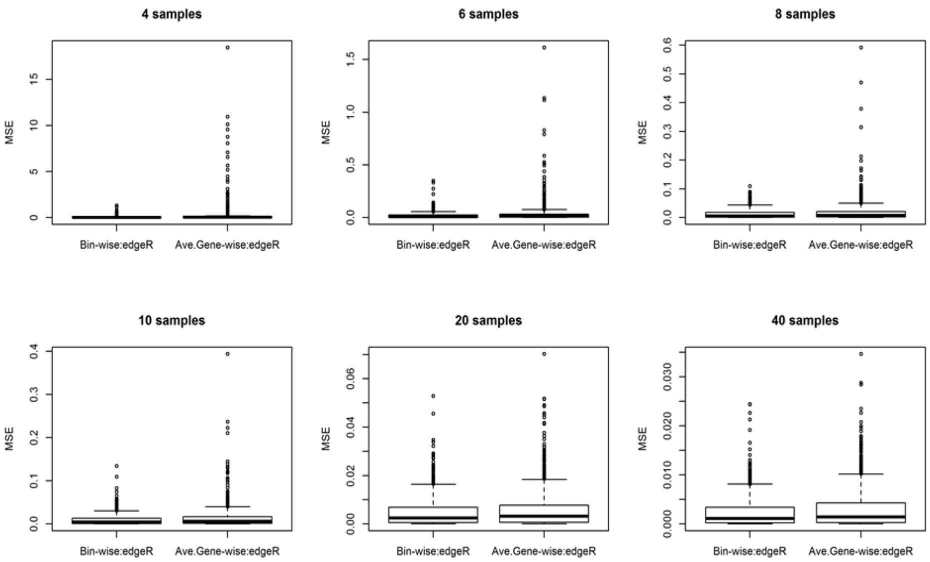

Fig. 2. Comparison of local estimation. The distribution of MSEs over 100 simulations under six different sample sizes. Each simulation is comprised of sampling of 10 transcripts with different mean and a constant ϕ=1/3. Bin-wise local estimation and average of 10 gene-wise estimates were compared. Comparison of methods: simulation study

3. Results and discussion

3.1. *Local binning improves estimation of small sample NB dispersion*

Estimation of dispersion parameter is a crucial step for differential expression analysis. Here, we compared bin-wise estimation and gene-wise estimation by studying an example of a single bin consisting of 10 hypothetical genes with the same dispersion value but with different means. We set the means of 10 genes from 1 to 10 and fixed the dispersion parameter ϕ=1. Using this hypothetical dataset, we compared two different estimation strategies (bin-wise estimation and average of gene-wise estimation) in terms of mean squared error (MSE). Since we set the ϕ=1, MSE is defined as,

$$MSE = \frac{1}{10}\sum_{i=1}^{10}\left(\hat{\phi}_i - 1\right)^2 \qquad (3)$$

To incorporate the effect of sample size in estimation performance, we varied the number of samples from 4 (2 in each condition) to 40 (20 in each condition). As can be seen in the Table 1 and Fig. 2, bin-wise estimates showed smaller MSEs. The distribution of MSEs indicated that bin-wise estimation provides more robust and accurate estimation of dispersion parameters in general (Fig. 2 and Table 1). Thus we used bin-wise local estimation in developing our proposed method.

Table 1. Comparison of bin-wise and gene-wise estimation.

The number of samples	MSE (Bin-wise estimation)	MSE (Average of gene-wise estimation)
4	0.0419	0.2092
6	0.0192	0.0351
8	0.0124	0.0177
10	0.0091	0.0137
20	0.0048	0.006
40	0.0024	0.0031

3.2. *Comparison of methods: simulation study*

The variance in NB model is a function of the mean and the dispersion parameter. Thus the performance of a statistical test largely depends on both parameters. To show that

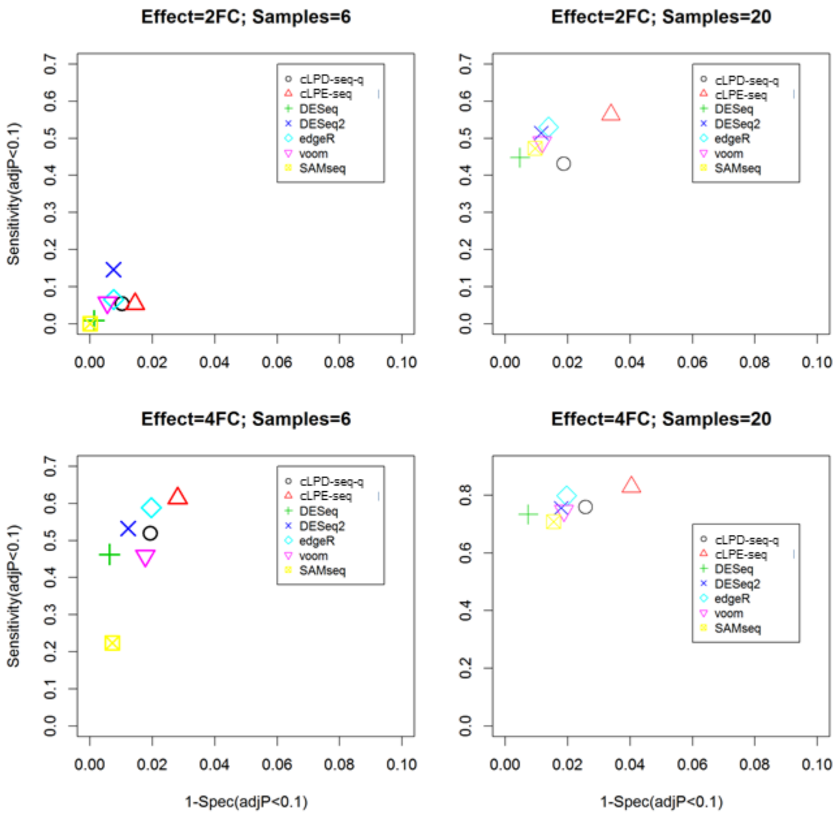

Fig. 3. Methods comparison using simulated data analysis. X- and Y-axis represent sensitivity and 1-specificity, respectively. Different fold change (FC) and sample sizes were plotted. Seven methods were applied and denoted by different shapes and colours. DE calls were made with the same criteria as was used in Love *et al* (BH-adjusted p-value < 0.1).

improvements in estimation of dispersion parameter can affect the performance of DE analysis, we repeated a subset of the simulation study of Love *et al*. [5].The simulation in Love *et al.* considered Fig. 3 shows the plot of sensitivity vs 1-specificity plot, where DE analysis was performed using a criterion of Benjamini-Hochberg adjusted P-value < 0.1 in accordance with Love *et al*. [5]. We compared our proposed methods, *cLPD-seq* (the equidistance binning approach) and *cLPD-seq-quantile* (the quantile binning approach), to five other competing methods.

The first situation considered an extreme case: with small effect size (2 fold change) and a small sample (3 per each group). In this situation, *DESeq2* performed the best, but all methods performed poorly with a seriously low power of less than 20%. Except for the extreme case above, *cLPD-seq* showed the highest power in all comparisons, as well as the highest 1-specificity (or equivalently false positive rate) less than or about 4% (red triangle in Fig 3). We see that using local-bin estimation with the equidistant mode results in the highest power on average in most of cases, with accessible false positive rates.

3.3. *Application to real RNA-Seq data*

We applied the proposed method *cLPD-seq* to the RNA-Seq datasets: Montgomery *et al.* [18] and Gilad *et al.* [19]. We did not use *cLPD-seq-quantile* to due to their poor performance in simulation studies. We compared the results of our proposed method only with *edgeR*, *DESeq*, *DESeq2*, and *voom*. *SAMseq* was excluded in comparison because of its lower power in small sample analysis.

To assess how well the methods control for the false positive rates, we artificially constructed two groups from the multiple samples with the same condition in Montgomery *et al*'s dataset. More specifically, we randomly selected 6 samples out of 27 male samples then performed DE analysis as if they were from two different groups. We repeated these

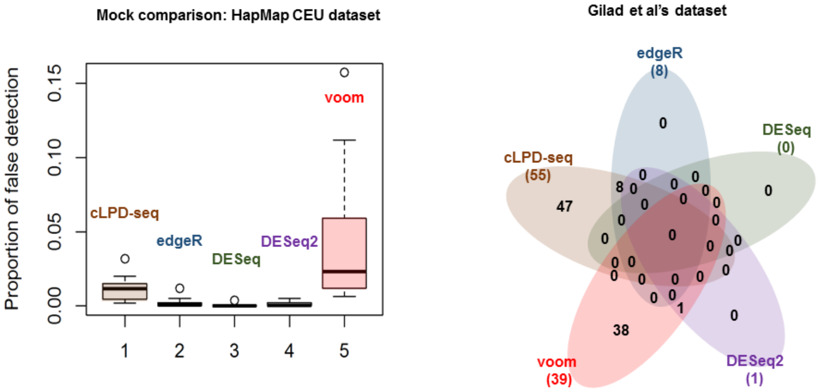

Fig. 4. Methods comparison using real data analysis. Performance of each method was considered using mock comparison of HapMap CEU samples in the same condition (left) and two-group comparison of sex-different samples (right). The number in parenthesis indicates the total number of DE transcript each method identified.

steps 100 times. Note that no significant detections were expected in these mock comparisons. All methods well conserved the proportion of false detection to less than 5%. *DESeq2* detected the least number of DE transcripts (average about 5). Though *cLPD-seq* showed a slightly higher proportion of false detection, an average of 1.2%, it is an acceptable rate (Fig. 4 left). *voom* showed the largest proportion of false detection and the largest variability. When we increased the sample size (sample size of 5), the results were similar (data not shown).

Gilad *et al*'s dataset [19] consists of liver RNA samples from three males and three females. A majority of the methods detected few or even no DE transcripts, except our proposed *cLPD-seq* method and *voom*. The number of total DEGs identified by *cLPD-seq* and *voom* were 55 and 39, respectively (Fig. 4). Interestingly there was no overlap between these two methods. To interpret the findings biologically, we performed gene set enrichment analysis using DAVID [23]. The number of DEGs identified by *edgeR*, *DESeq* and *DESeq2* was 9, 0 and 1, respectively, and no gene set was enriched with these lists. Gene ontology (GO) analysis for the 55 DEGs identified by the proposed *cLPD-seq* reported six significant biological processes (FDR < 10%). Among enriched terms, female characteristic-related terms, i.e. response to estradiol stimulus (FDR = 0.6%) and response to estrogen stimulus (FDR = 4.5%) were identified. On the other hand, the GO analysis for the DEGs identified by *voom*, did not provide any significant biological processes. This simple application of male and female dataset demonstrates that our *cLPD-seq* provides more reasonable result than *voom*.

It is also interesting to evaluate how well the methods perform on small subsamples. It is reasonable to assume that the larger dataset gives more reliable results, leading them to be used as an underlying truth. Based on this idea, we performed DE analyses with subsamples taken from the original total samples and compared the results to the underlying truth. This so-called 'overlapping' analysis can be conducted with either all significant genes (likely to be thousands of) or a part of them (top *k* significant genes). For the latter, a ranking-invariant constraint is incorporated into the comparison by assuming that the top ranked genes should be more enriched in any robust methods. Also it is much easier to interpret the result and to carry out a follow-up study with a small number of genes. We performed the both analyses and repeated 10 times to evaluate whether the results from different subsamples agree with each other.

The results with all significant genes and top 200 significant genes are respectively shown in left and right of Fig 5. The overall performance with all DEGs were comparable for five methods. The number of DEGs, as well as the proportion of overlaps, identified by *DESeq*, and *DESeq2* were slightly lower compared to *cLPDseq*, *edgeR* and *voom* with all DEGs (left in Fig 5). The proposed method, *cLPDseq*, *edgeR*, and *voom* showed similarly performed better than the others regardless of the number of samples, but varied with different subsamples except *edgeR* with extremely small samples. In the case of sample size = 2, the average proportions (standard deviation) of *cLPDseq, edgeR, DESeq, DESeq2,* and *voom* are 0.762 (0.0662), 0.769 (0.02362), 0.663 (0.08888), 0.575 (0.08147), and 0.754 (0.08307) in order. Unlike the result with all DEGs, the overlap proportions with the top

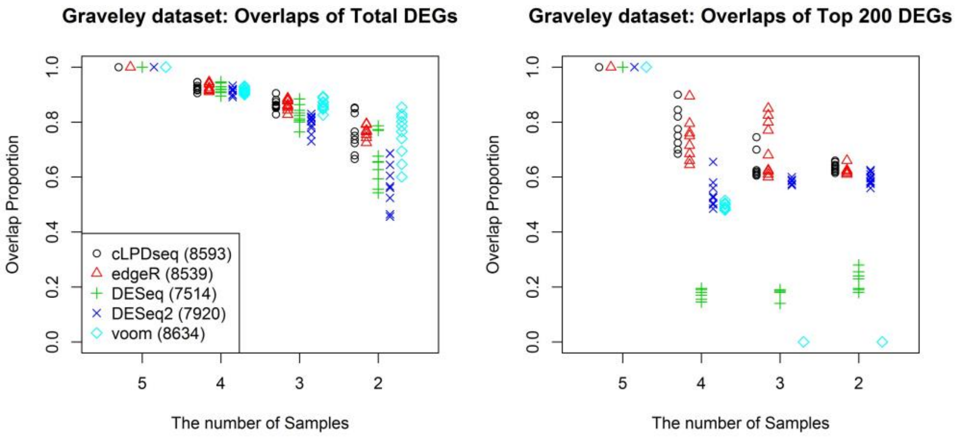

Fig. 5. Reproducibility and stability of methods. Overlap proportion with all DEGs (left) and top 200 DEGs (right) are shown. X and Y axis represent the number of samples in each condition and the proportion of DEGs overlapped with the result with total samples, respectively. The analysis was repeated 10 times with randomly selected subsamples for a specific sample size per each method, distinguished by colour. The numbers in parenthesis represent the number of DEGs identified with the total samples.

200 genes were quite distinct. With decreased number of samples, *DESeq* and *voom* showed poor overlap proportion. The overlap proportions of *cLPDseq* and *edgeR* were superior to others in both analyses, showing robust performance of these methods. Note that the number of incorporated genes is increased, the overlaps of each method is also increased.

4. Conclusion

To assess the significance of expression changes between different classes, estimation of dispersion parameter is critical. For RNA-Seq data, the models based on NB distribution assumption have been well studied and their superior performance to other approaches have been successfully demonstrated. Nonetheless, for data with a small number of replication, the lack of power was a limitation. (Nonetheless, data with a small number of replication always suffer from the lack of power.) We have integrated a local pooling idea into a common dispersion estimation idea by using conditional likelihood estimation for local pooled dispersion parameters in RNA-Seq data.

In our analyses using both simulation and real data application, it seemed as though estimation using local bins worked well in practice, adjusting the dispersion parameter estimation more closely and robustly to the true value and thus improving DE detection power. Note that the method showed robust results (similar DE performance) with different binning approach and different number of bins (data not shown). However, there are several issues requiring some further investigation: choosing an optimal number of bins and a proper smoothing curve. For future studies, these issues should be dealt more rigorously. A reproducible code, datasets, and supplementary figures will be found in website (http://bibs.snu.ac.kr/software/cLPDseq)

Acknowledgments

This work is supported by the Bio-Synergy Research Project (2013M3A9C4078158) of the Ministry of Science, ICT and Future Planning through the NRF.

References

1. Soneson, C. and M. Delorenzi, *A comparison of methods for differential expression analysis of RNA-seq data.* Bmc Bioinformatics, 2013. **14**.
2. Robinson, M.D. and G.K. Smyth, *Moderated statistical tests for assessing differences in tag abundance.* Bioinformatics, 2007. **23**(21): p. 2881-7.
3. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-140.
4. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.
5. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 2014. **15**(12).
6. Li, J. and R. Tibshirani, *Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data.* Stat Methods Med Res, 2013. **22**(5): p. 519-36.
7. Law, C.W., et al., *voom: Precision weights unlock linear model analysis tools for RNA-seq read counts.* Genome Biol, 2014. **15**(2): p. R29.
8. Hardcastle, T.J. and K.A. Kelly, *baySeq: empirical Bayesian methods for identifying differential expression in sequence count data.* BMC Bioinformatics, 2010. **11**: p. 422.
9. Van De Wiel, M.A., et al., *Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors.* Biostatistics, 2013. **14**(1): p. 113-28.
10. Tarazona, S., et al., *Differential expression in RNA-seq: a matter of depth.* Genome Res, 2011. **21**(12): p. 2213-23.
11. Rapaport, F., et al., *Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data.* Genome Biology, 2013. **14**(9).
12. Seyednasrollah, F., A. Laiho, and L.L. Elo, *Comparison of software packages for detecting differential expression in RNA-seq studies.* Briefings in Bioinformatics, 2015. **16**(1): p. 59-70.
13. Jain, N., et al., *Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays.* Bioinformatics, 2003. **19**(15): p. 1945-1951.
14. Robinson, M.D. and G.K. Smyth, *Small-sample estimation of negative binomial dispersion, with applications to SAGE data.* Biostatistics, 2008. **9**(2): p. 321-32.
15. Lonnstedt, I. and T. Speed, *Replicated microarray data.* Statistica Sinica, 2002. **12**(1): p. 31-46.

16. Frazee, A.C., B. Langmead, and J.T. Leek, *ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets.* BMC Bioinformatics, 2011. **12**: p. 449.
17. Pickrell, J.K., et al., *Understanding mechanisms underlying human gene expression variation with RNA sequencing.* Nature, 2010. **464**(7289): p. 768-72.
18. Montgomery, S.B., et al., *Transcriptome genetics using second generation sequencing in a Caucasian population.* Nature, 2010. **464**(7289): p. 773-7.
19. Blekhman, R., et al., *Sex-specific and lineage-specific alternative splicing in primates.* Genome Res, 2010. **20**(2): p. 180-9.
20. Graveley, B.R., et al., *The developmental transcriptome of Drosophila melanogaster.* Nature, 2011. **471**(7339): p. 473-9.
21. Robinson, M.D. and A. Oshlack, *A scaling normalization method for differential expression analysis of RNA-seq data.* Genome Biol, 2010. **11**(3): p. R25.
22. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.
23. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature Protocols, 2009. **4**(1): p. 44-57.

Jungsoo Gim received his Ph.D. degree in Bioinformatics from Seoul National University, South Korea in 2014. From 2014 to 2015, he worked as a post doctor at the Department of Statistics, Seoul National University. He is currently a research assistant professor at Institute of Health and Environment, Seoul National University since 2015. His research interests include developing statistical methods for OMICS data analysis in genome-wide scale and disease prediction model.

Sungho Won received his Ph.D. degree in Epidemiology and Biostatistics from Case Western Reserve University, USA 2008. He was a research fellow and a research associate in Department of Biostatistics at Harvard University from July 2008 to June 2009, and from July 2009 to Aug 2009, respectively. From 2010 to 2014, he worked as an assistant professor in Department of Applied Statistics from Chung-Ang University, South Korea. Since 2014, he moved to Department of Public Health Science, Seoul National University and he is currently an associate professor. His research interests include biostatistics, bioinformatics and genetic epidemiology.

Taesung Park received his Ph.D. degree in Biostatistics from University of Michigan, USA in 1990. From September 1999 to September 2001, he worked as an associate professor in Department of Statistics from Seoul National University, Korea. From October 2001 he worked as a professor in Department of statistics from Seoul National University, Korea. He is a director of Creative Research Lab for Bioinformatics and Biostatistics, Seoul National University. His research areas include genome-wide association studies, statistical genetics and longitudinal data analysis