論文の内容の要旨

応用生命工学専攻
平成 25 年度博士入学
氏名　　湯敏
指導教員名　清水謙多郎

論文題目

A comprehensive evaluation of methods for differential expression analysis on multi-group RNA-seq count data

（RNA−seqの多群間比較用カウントデータに基づく発現変動解析手法の評価）

Introduction

A major application of high-throughput sequencing (HTS) is to measure RNA transcript expression (so-called RNA-seq). While microarrays are limited to the detection of known transcripts and to discriminate between transcript variants, RNA-seq can detect all transcripts in the cell in light of their sequence, structure, and expression levels. Identifying differentially expressed genes or transcripts (DEGs) in different groups or conditions is one important goal for RNA-seq. The differential expression (DE) analysis typically starts with a count matrix, the first row denotes the sequencing sample names while the first column denotes the gene (or transcript or genomic loci) names, and each of the other cells is filled with the number of reads mapped to the gene in the sample (i.e., counts).

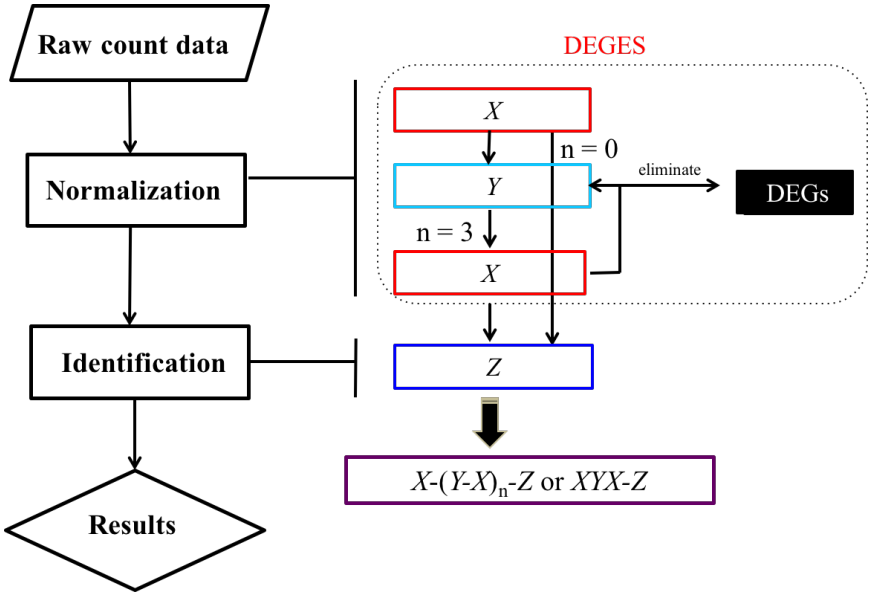
Up to now, there are a number of methods for detecting DEGs in RNA-seq count data. Many of these methods can perform not only two-group comparison but also more complex experimental designs such as multi-group (>2) and multi-factored ones. However, their evaluations have been limited to two-group comparisons (e.g., two cellular conditions or phenotypes). Researchers also wish to know which method outperforms the others when performing multi-group comparisons. In general, the procedure of detecting DEGs is constructed as a pipeline (called DE pipeline) of two kinds of steps: data normalization and DEG identification. The propose of this study is to compare DE pipelines for multi-group data, focusing on the three-group comparison, and to provide the guidelines.

Methods

We investigated a total of 12 pipelines available in nine R packages: TCC (ver. 1.7.15), edgeR (ver. 3.8.5), DESeq (ver. 1.18.0), DESeq2 (ver. 1.6.3), voom in limma (ver. 3.22.1), SAMseq in samr (ver. 2.0), PoissonSeq (ver. 1.1.2), baySeq (ver. 2.0.50), and EBSeq (ver. 1.6.0). Of these packages, TCC was developed by our group. Different from the other packages, TCC implements a multi-step normalization strategy (called DEGES) that internally uses functions provided by other representative packages (edgeR, DESeq, and DESeq2).

We denote the data normalization and DEG identification steps of the DE pipelines as *X* and *Y*, respectively. Therefore, the whole procedure of RNA-seq count data can be described as *X*-*Y* pipeline. In the most of the publicly available R packages, individual R package has its own methods for the elements of *X*-*Y* pipeline. For example, when comparing multi-group data in edgeR, the default normalization method is TMM (Trimmed Means of M values), and the default DEG identification method is the likelihood ratio test based on generalized linear models (GLM) whose error structure follows the negative binomial distribution. One of the models corresponds to alternative hypothesis and the other corresponds to null hypothesis. In this study, we termed the default DE pipelines *X*-*Y* for edgeR, DESeq, and DESeq2 "edgeR-edgeR (or *E*-*E*)", "DESeq-DESeq (or *D*-*D*)", and "DESeq2-DESeq2 (*S*-*S*)", respectively.

The DE pipeline implemented in TCC can be described as *X*-(*Y*-*X*)$_n$-*Y* or *XYX*-*Y*. The DEGES normalization in TCC corresponds to *X*-(*Y*-*X*)$_n$. The key concept for the normalization strategy is to alleviate the negative effect of potential DEGs before calculating the normalization factors in step 3. The DEG elimination strategy with $n \geq 2$ corresponds to the iterative DEGES or iDEGES in TCC. Recommendations made in the original TCC paper were (i) the iDEGES/edgeR-edgeR pipeline on two-group data with replicates and (ii) the iDEGES/DESeq-DESeq pipeline on two-group data without replicates. In this study, we abbreviated these pipelines to *EEE*-*E* and *DDD*-*D* for convenience. This is mainly because (1) users can select different DEG identification methods *Y* for steps 2 and 4 and (2) we compare several possible combinations such as *EDE*-*S* for the edgeR-(DESeq-edgeR)$_n$-DESeq2 pipeline. Since *n* is usually fixed as $n = 3$, the DEGES-based DE pipeline can also be described as *XYX*-*Z*, where *Y* and *Z* are the same or different DEG identification methods (Figure 1). For simplicity, three pipelines (i.e., *EEE*-*E*, *DDD*-*D*, and *SSS*-*S*) in TCC were mainly compared.



We evaluated those pipelines on the basis of both simulation data and real count data. To evaluate the multi-group count data as simply as possible, we focused on the three-group comparison (i.e., G1 vs. G2 vs. G3) with equal number of biological replicates (i.e., 1, 3, 6

and 9 replicates per group; Nrep = 1, 3, 6, and 9). The simulation conditions are: the total number of genes is 10,000 (Ngene = 10000), 5 or 25% of the genes are DEGs (P_{DEG} = 5 or 25%), the levels of DE are four-fold in individual groups, and a total of seven different proportions of DEGs up-regulated in individual groups (P_{G1}, P_{G2}, P_{G3}), from unbiased pattern (1/3, 1/3, 1/3) to strongly biased pattern (0.8, 0.1, 0.1) (Table 1). The expression levels of up-regulated genes are obviously higher in the designated group than in the other two groups. The simulation analyses were performed using the functions provided in TCC. We also analyzed a real three-group data with six biological replicates from three species (human, chimpanzee and rhesus macaques). The publicly available count matrix consisting of 20,689 genes × 18 samples were mainly evaluated on the basis of the overall similarity and reproducibility of DE results produced from the 12 DE pipelines.

Results and discussion

The comparisons were mainly performed using the AUC values (area under the curve) to evaluate both sensitivity and specificity of the pipelines simultaneously. Table 1 shows the average AUC values of 100 trials of pipelines for four different proportions of DEGs with P_{DEG} = 25%. When comparing the results of count data with replicates (Nrep = 3; Table 1a), the AUC values for the *EEE-E* pipeline were the highest and similar across the four different proportions of DEGs up-regulated in individual groups (P_{G1}, P_{G2}, P_{G3}). The edgeR (i.e., *E-E*) performed the second best overall. We also observed this trend under increased numbers of replicates (i.e., Nrep = 6 and 9; data not shown).

Table 1 - Average AUC values (%) for three-group simulation data with P_{DEG} = 25%

PG1	33%	50%	60%	80%	33%	50%	60%	80%
PG2	33%	40%	30%	10%	33%	40%	30%	10%
PG3	33%	10%	10%	10%	33%	10%	10%	10%
	(a) with replicates (Nrep = 3)				(b) without replicates (Nrep = 1)			
EEE-E (TCC)	91.47	**91.45**	**91.43**	**91.37**	77.15	76.78	76.88	75.48
DDD-D (TCC)	90.77	90.72	90.68	90.57	81.51	81.26	81.15	79.98
SSS-S (TCC)	88.13	88.13	88.12	88.06	**82.46**	**82.08**	**82.18**	**80.97**
E-E (edgeR)	**91.47**	91.18	90.98	89.97	77.15	76.76	76.87	75.36
edgeR_robust	90.89	90.57	90.34	89.27	-	-	-	-
D-D (DESeq)	90.77	90.37	90.15	89.04	81.53	81.23	81.12	79.84
S-S (DESeq2)	88.12	87.62	87.36	85.92	82.46	82.01	82.16	80.73
voom	87.08	86.52	86.18	84.56	-	-	-	-
SAMseq	84.95	84.82	84.75	84.63	-	-	-	-
PoissonSeq	87.22	87.14	87.11	86.97	-	-	-	-
baySeq	90.34	90.07	89.83	88.86	-	-	-	-
EBSeq	85.82	85.49	85.30	84.02	-	-	-	-

Unlike multi-group count data with replicates, there are few R packages (including TCC, edgeR, DESeq, and DESeq2) that can manipulate data without replicates. When comparing the results of count data without replicates (Nrep = 1; Table 1b), the AUC values for the *SSS-S* pipeline were the highest. When three original non-DEGES-based pipelines *X-Z* are compared, DESeq2 (i.e., *S-S*) performed the best, followed by DESeq and edgeR. This order was exactly

opposite from the results in Table 1a. We also found that choosing *Z* for the DEGES-based pipeline has more impact on the accuracy of DE result than choosing *Y*, despite of the number of replicates (data not shown).

With respect to the real data analysis, we found that seven GLM-based DE pipelines (*SSS-S*, *S-S*, edgeR_robust, *EEE-E*, *E-E*, *DDD-D*, *D-D*) showed similar ranked gene lists. It should be noted that the overall similarity does not necessarily indicate the superiority of the seven pipelines over the other pipelines such as EBSeq. Indeed, EBSeq consistently outperformed the others when analyzing simulation data with Nrep = 9 and P_{DEG} = 5%. EBSeq (and baySeq) employs an empirical Bayesian framework that outputs the posterior probabilities for each of the predefined possible expression patterns to each gene. This is probably the main reason for EBSeq and baySeq to have lower similarity than the other GLM-based pipelines, indicating that the lower similarity between different algorithms do not matter. We found that different pipelines could produce considerably different numbers of DEGs from 3,832 (DESeq) to 9,453 (SAMseq). With respect to the reproducibility (numbers of common genes between the different subsets), we observed high reproducibility for three pipelines (*EEE-E*, *E-E* and edgeR_robust) and low reproducibility for two pipelines (SAMseq and EBSeq).

Conclusion

To our knowledge, this work is the first comprehensive study on multi-group count data. Our main findings can be summarized as follows: First, the idea of DEGES implemented in TCC can be applied to multi-group data and the DEGES-based pipelines can be performed easily in TCC package. Second, the choice of DEG identification method *Z* in the DEGES-based pipeline *XYX-Z* is critical for obtaining good DE results. Third, to analyze three-group data with replicates, we recommend using *EEE-E*. Fourth, to analyze three-group data without replicates, we recommend using either *SSS-S*.

Reference

Tang M, Sun J, Shimizu K, Kadota K. Evaluation of methods for differential expression analysis on multi-group RNA-seq count data. *BMC Bioinformatics*, 2015, **16**: 361.