Accurate differential gene expression analysis for RNA-Seq data without replicates

Sahar Al Seesi[1], Yvette Temate Tiagueu[2], Alex Zelikovsky[2], and Ion Măndoiu[1]

[1] Computer Science & Engineering Department, University of Connecticut
{sahar,ion}@engr.uconn.edu
[2] Computer Science Department, Georgia State University
ytematetiagueu1@student.gsu.edu, alexz@cs.gsu.edu

Abstract. Despite the fact that many RNA-Seq experiments do not use biological replicates, most of the existing differential gene expression analysis for RNA-Seq data are designed to work with replicates. We present a novel method for differential gene expression analysis based on bootstrapping. We also discuss the use of different normalization methods with Fisher's exact test. We compare the methods we present, with existing methods, on real RNA-Seq benchmarking data. Our comparison shows that bootstrapping outperforms other methods in most cases.

1 Introduction

RNA-Seq is the new standard for the analysis of differential gene expression [9] [8] [14]. For this purpose, RNA-Seq produces gene expression profiles with much smaller technical variance [3] than traditional microarray technologies. However, simply using the raw fold change of the expression levels of a gene across two samples, as a measure of differential expression, can still be unreliable, because it does not account for the uncertainty in the gene expression estimation. Therefore, the need for using statistical methods arises. Despite the fact that most RNA-Seq data do not have biological replicates [2], most reliable differential expression analysis approaches are designed for data with replicates. EdgeR[11] and DESeq [1], are both statistical packages which work with replicates, and compute exact test based on negative binomial distribution. Recently, Feng el al. presented GFOLD [5], a generalized fold change algorithm which produces biologically meaningful rankings of differentially expressed genes from RNA-Seq data. They show that GFOLD outperforms methods designed to work with replicates, when used for single replicate datasets.

In this work, we present a novel method for differential expression analysis based on bootstrapping, and a comparison for a number of methods including different normalization methods that can be used with Fisher's exact test. The comparison includes GFOLD as well as cuffdiff, which is part of the widely used cufflinks package [12]. We assess the accuracy of these methods using the RNA-Seq data generated for MAQC [7] samples using two different technologies (Illumina and ION Torrent). The comparisons are done for different fold change thresholds.

2 Methods

Here we discuss two statistical methods for differential gene expression analysis from RNA-Seq data. The first method is a novel method that uses bootstrapping to calculate the support of a fold change for the expression of a gene across two samples. The second is Fisher's exact test which was suggested by Bullard et al. [3]. We we look at different normalization methods to be used with Fisher's exact test, including using synthetic RNA spike-in controls.

2.1 Bootstrapping

We use bootstrapping to calculate the fold change for the expression of a gene between two datasets. Given two RNA-Seq read sets, a and b, we generate 200 random samples with replacement for the reads in a, and similarly for b, where the size of each sample equals the minimum of the total number of reads in a and b. In each resampling iteration, we use the alignments of the set of selected reads to calculate gene expression levels. For efficiency, we map the reads in a and b once, and we extract the alignments of the sampled reads in each iteration. For gene expression estimation, we use IsoEM [10] to calculate the Fragment per kilobase of gene length per Million reads (FPKM). IsoEM is also run on the complete sets of read alignments for a and b and the estimated FPKM values are used to determine the direction of over expression D_g, if any, for each gene, for a specific fold change threshold. To make a differential expression call of a gene g, 200 fold changes are calculated for g by randomly select $FPKM_{a_i}(g)$ and $FPKM_{b_i}(g)$ for $i = 1..200$ from the FPKM estimates calculated in the bootstrapping iterations. For a given fold change, x, we calculate the the percentage p of fold changes of $g \geq x$, where the direction of over expression of these fold cahnges agrees with D_g. A gene g is called differentially expressed if $p \geq 50\%$.

2.2 Fisher's exact test

Fisher's exact test is a statistical significance test for categorical data which measures the association between two variables. The data is classified in a 2x2 contingency table according to the two variables of interest. We use Fisher's exact test to measure the statistical significance of change in gene expressions between two samples a and b by setting the two values in the first row of table to the estimated number of reads mapped per kilobase of gene length (calculated from IsoEM estimated FPKM values) in samples a and b. The values in the second row of the contingency table depend on the normalization method used. We compare three normalization methods. The first one is total read normalization, where the total number of mapped reads in samples a and b are used in the second row. The second is normalization by a housekeeping gene. In this case, the estimated number of reads mapped per kilobase of housekeeping gene length in each sample is used. We also test normalization by External RNA Controls Consortium (ERCC) RNA spike-in controls. FPKMs of ERCCs

are aggregated together (similar to aggregating the FPKMs of different transcripts of a gene), and the estimated number of reads mapped per kilobase of ERCC are calculated from the resulting FPKM value and used for normalization. The calculated p-value which measures the significance of deviation from the null-hypothesis, namely the gene being not differentially expressed, is exactly measured by calculating the hypergeometric probability of the numbers given in the contingency table or more extreme differences, while keeping the marginal sums in the contingency table unchanged. We adjust the resulting p-values for the set of genes being tested for 5% false discovery rate (FDR).

3 Experimental Results

We conducted experiments on RNA-Seq data generated from two commercially available reference RNA samples that have been well-characterized by quantitative real time PCR (qRT-PCR) as part of the MicroArray Quality Control Consortium (MAQC) [7]; namely an Ambion Human Brain Reference RNA, Catalog # 6050), henceforth referred to as HBRR and a Stratagene Universal Human Reference RNA (Catalog # 740000), henceforth referred to as UHRR. To assess accuracy, DE calls obtained from RNA-Seq data were compared against those obtained from TaqMan qRT-PCR measurements (GEO accession GPL4097) collected as part of the MAQC project. Each TaqMan Assay was run in four replicates for each measured gene. POLR2A (ENSEMBL id ENSG00000181222) was chosen as the reference gene and each replicate CT was subtracted from the average POLR2A CT to give the log2 difference (delta CT). For delta CT calculations, a CT value of 35 was used for any replicate that had CT >35. The normalized expression value of a gene g would be: $2^{(CT_{POLR2A} - CT_g)}$. We filtered out genes that: (1) were not detected present in two or more replicates in each samples or (2) had a standard deviation higher than 25% for the four TaqMan values in each of the two samples. Of the resulting subset, we used in the comparison genes whose TaqMan probe ids mapped to Ensemble gene ids (686 genes).

		True	
Predicted	Over-Expressed (TOE)	Non-Differential (TND)	Under-Expressed (TUE)
Over-Expressed (POE)	TPOE		
Non-Differential (PND)		TPND	
Under-Expressed (PUE)			TPUE

Table 1: Confusion Matrix for Differential Expression

For the ground truth, a gene was considered differentially expressed if the fold change in the average normalized TaqMan expression levels between the two samples is greater than a set threshold with the p-value for an unpaired two-tailed T-test (adjusted for 5% FDR) is less than 0.05. We ran the experiment

for fold change thresholds of 1, 1.5, and 2. For each RNA-Seq differential expression method being evaluated, genes were classified according to the differential expression confusion matrix detailed in Table 1. Methods were assessed using sensitivity, positive predictive value (PPV), F-score, which is the harmonic mean of sensitivity and PPV, and accuracy, defined as follows:

$$\begin{aligned} Sensitivity &= \frac{(TPOE+TPUE)}{(TOE+TUE)} & PPV &= \frac{(TPOE+TPUE)}{(POE+PUE)} \\ Accuracy &= \frac{(TPOE+TPND+TPUE)}{(TOE+TND+TUE)} & F\text{-}score &= 2 \times \frac{TPR \times SPC}{TPR+SPC} \end{aligned}$$

We compared different methods on Illumina and ION Torrent datasets. For Illumina, we downloaded HBBR SRX003926 and UHHR SRX003927 datasets [13] from NCBI Short Read Archive. Reads were mapped to hg19 Ensembl 63 transcript library, using bowtie v0.12.7.0 [6]. Mapping resulted in 6.8 and 5.8 million mapped reads, for HBBR and UHHR datasets respectively. We used this dataset to compare cuffdiff, GFOLD, and Fisher's exact test, using both total and housekeeping gene (POLR2A) normalization, and bootstrapping. We used cuffdiff v2.0.1 with default parameters and GFOLD v1.0.7 with default parameters and fold change significant cuttof of 0.05. Number of mapped reads per kilobase of gene length used in Fisher's exact test calculation are based on IsoEM FPKMs.

For testing on ION-Torrent, we merged five UHRR (DID-144-283, GOG-140-284, POZ-125-268, POZ-126-269, and POZ-127-270) together and five HBRR together (DID-143-282, GOG-139-281, LUC-140-265, LUC-141-267, and POZ-124-266). Reads were mapped to hg19 Ensembl 64 transcript library using tmap v2.3.2. The number of mapped reads were 5.51 and 6.6 million reads, for HBRR and UHHR respectively. For this dataset, we compared three different normalization methods for Fisher's exact test; namely total normalization, housekeeping gene (POLR2A) normalization, and normalization using ERCCs [4] which were spiked in these RNA samples. Bootstrapping and GFOLD were included in the comparison. We did not include cuffdiff in this comparison due the big gap in performance it showed, compared to other methods on the Illumina dataset.

Table 2 shows the results obtained from the Illumina dataset from fold change 1, 1.5 and 2. Table 3 shows the results obtained from the ION Torrent dataset for the same fold changes. The best performing method for each statistic, within a fold change, is highlighted in bold. Comparisons show that bootstrapping outperforms other methods in most cases, specially at a lower fold change threshold. Fisher exact tests present comparable results to other methods. Total count notmalization gives the beat results for Fisher's exact test at fold change 1; however this changes in favor of housekeeping gene normalization, compared to both total and ERCC normalization.

Acknowledgments. This work has been partially supported by Collaborative Research Grant AG110891 from Life Technologies, awards IIS-0546457, IIS-0916401, and IIS-0916948 from NSF, Agriculture and Food Research Initiative

Fold Change	Method	Accuracy %	Sensitivity %	PPV %	F-Score %
1	FishersTotal	70.41%	70.79%	91.24%	79.72%
	FishersHousekeeping	65.60%	65.22%	95.05%	77.36%
	GFOLD	78.13%	80.06%	92.67%	**85.90%**
	Cuffdiff	11.37%	6.96%	**100.00%**	13.01%
	Bootstrapping	**82.22%**	**87.17%**	83.06%	85.07%
1.5	FishersTotal	74.05%	78.20%	84.85%	81.39%
	FishersHousekeeping	76.68%	73.61%	93.67%	82.44%
	GFOLD	79.15%	79.35%	90.41%	84.52%
	Cuffdiff	28.43%	8.60%	**100.00%**	15.85%
	Bootstrapping	**79.74%**	**87.76%**	84.38%	**86.04%**
2	FishersTotal	78.43%	81.86%	82.44%	82.15%
	FishersHousekeeping	81.20%	80.00%	88.21%	83.90%
	GFOLD	**82.94%**	78.84%	92.37%	**85.07%**
	Cuffdiff	40.96%	10.47%	**100.00%**	18.95%
	Bootstrapping	80.76%	**86.74%**	82.71%	84.68%

Table 2: Accuracy, sensitivity, PPV and F-Score in % for Illumina dataset and Fold-Change = 1, 1.5, and 2

Fold Change	Method	Accuracy %	Sensitivity %	PPV %	F-Score %
1	FisherTotal	71.68%	72.76%	90.56%	80.69%
	FisherHousekeeping	67.15%	66.87%	**94.74%**	78.40%
	FisherERCC	71.39%	72.45%	88.97%	79.86%
	GFOLD	75.77%	77.55%	90.43%	83.50%
	Bootstrapping	**82.19%**	**86.84%**	82.87%	**84.81%**
1.5	FisherTotal	74.16%	78.39%	85.06%	81.59%
	FisherHousekeeping	76.06%	73.23%	**92.96%**	81.93%
	FisherERCC	74.31%	78.59%	85.45%	81.87%
	GFOLD	75.47%	77.63%	87.88%	82.44%
	Bootstrapping	**77.81%**	**85.28%**	83.83%	**84.55%**
2	FisherTotal	79.71%	83.02%	84.00%	83.51%
	FisherHousekeeping	**81.75%**	80.70%	88.75%	84.53%
	FisherERCC	79.42%	82.56%	84.12%	83.33%
	GFOLD	80.58%	76.74%	**90.66%**	83.12%
	Bootstrapping	80.88%	**86.05%**	83.33%	**84.67%**

Table 3: Accuracy, sensitivity, PPV and F-Score in % for Ion Torrent dataset and Fold-Change = 1, 1.5, and 2

Competitive Grant no. 201167016- 30331 from the USDA National Institute of Food and Agriculture, and the Molecular Basis of Disease Area of Focus Georgia State University.

References

1. Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome Biol 11(10), R106 (2010)
2. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Muertter, R.N., Edgar, R.: NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acid Research 37, D885D890 (2011)
3. Bullard, J., Purdom, E., Hansen, K., Dudoit, S.: Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11(1), 94 (2010)
4. External RNA Controls Consortium: Proposed methods for testing and selecting the ERCC external RNA controls. BMC Genomics 6, 150 (2005)
5. Feng, J., Meyer, C.A., Wang, Q., Liu, J.S., Liu, X.S., Zhang, Y.: GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-Seq data. Bioinformatics 28(21), 2782–2788 (2012)
6. Langmead, B., Trapnell, C., Pop, M., Salzberg, S.: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10(3), R25 (2009)
7. MAQC Consortium: The Microarray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nature Biotechnology 24(9), 1151–1161 (Sep 2006)
8. Morozova, O., Hirst, M., Marra, M.A.: Applications of new sequencing technologies for transcriptome analysis. Annual review of genomics and human genetics 10, 135–151 (2009)
9. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B.: Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods 5(7), 621–628 (2008)
10. Nicolae, M., Mangul, S., Mandoiu, I.I., Zelikovsky, A.: Estimation of alternative splicing isoform frequencies from RNA-Seq data. In: Moulton, V., Singh, M. (eds.) Proc. WABI. Lecture Notes in Computer Science, vol. 6293, pp. 202–214. Springer (2010)
11. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1), 139–140 (2010)
12. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L.: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28(5), 511–515 (2010)
13. Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth4, G.P., Burge1, C.B.: Alternative isoform regulation in human tissue transcriptomes. Nature 457(7221), 470–476 (2008)
14. Wang, Z., Gerstein, M., Snyder, M.: RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10(1), 57–63 (2009)