BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 21 2012, pages 2782–2788
doi:10.1093/bioinformatics/bts515

Gene expression

Advance Access publication August 24, 2012

GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data

Jianxing Feng[1,2], Clifford A. Meyer[3], Qian Wang[1], Jun S. Liu[4], X. Shirley Liu[3,*] and Yong Zhang[1,*]

[1]Department of Bioinformatics, School of Life sciences and Technology, Tongji University, 1239 Siping Road, Shanghai 20092, China, [2]Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, 220 Handan Road, Shanghai, 200433, China, [3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard School of Public Health, Harvard University and [4]Department of Statistics, Harvard University, Science Center 715, 1 Oxford Street, Cambridge, MA 02138, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: RNA-seq has been widely used in transcriptome analysis to effectively measure gene expression levels. Although sequencing costs are rapidly decreasing, almost 70% of all the human RNA-seq samples in the gene expression omnibus do not have biological replicates and more unreplicated RNA-seq data were published than replicated RNA-seq data in 2011. Despite the large amount of single replicate studies, there is currently no satisfactory method for detecting differentially expressed genes when only a single biological replicate is available.

Results: We present the GFOLD (generalized fold change) algorithm to produce biologically meaningful rankings of differentially expressed genes from RNA-seq data. GFOLD assigns reliable statistics for expression changes based on the posterior distribution of log fold change. In this way, GFOLD overcomes the shortcomings of *P*-value and fold change calculated by existing RNA-seq analysis methods and gives more stable and biological meaningful gene rankings when only a single biological replicate is available.

Availability: The open source C/C++ program is available at http://www.tongji.edu.cn/~zhanglab/GFOLD/index.html

Contact: xsliu@jimmy.harvard.edu or yzhang@tongji.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 9, 2012; revised on August 13, 2012; accepted on August 14, 2012

1 INTRODUCTION

In RNA-seq, high-throughput sequencing is applied to transcriptome analysis to effectively measure gene expression levels, identify alternative splicing variants and reconstruct novel or fusion transcripts. Since the first publications in a series of studies in 2008 (Cloonan *et al*., 2008; Lister *et al*., 2008; Marioni *et al*., 2008; Morin *et al*., 2008; Mortazavi *et al*., 2008; Nagalakshmi *et al*., 2008; Wilhelm *et al*., 2008), RNA-seq has quickly gained popularity for use in transcriptome analysis (Haas and Zody, 2010; Morozova *et al*., 2009; Wall *et al*., 2009; Wang *et al*., 2009).

One fundamental use of transcriptome analysis is to measure the level of gene expression and to identify genes that are differentially expressed between conditions. For this purpose, RNA-seq produces gene expression profiles with much smaller technical variance (Bullard *et al*., 2010) than traditional microarray technologies. Specifically, in a typical RNA-seq experiment, millions of short reads are sampled from expressed transcripts and the expression level of a gene is then measured by the number of reads mapped back to this gene.

The variance in the read count of a gene may be decomposed into the variance due to the random sampling of reads and the variance due to other sources of variation including technical and biological noise. The variance due to read sampling can be closely approximated by a Poisson distribution (Jiang and Wong, 2009), which serves as the lower bound on the overall variance. When biological variance is taken into consideration, the number of reads mapped to a gene should resemble an over-dispersed Poisson distribution. A natural distribution for modeling over-dispersed count data is the negative binomial (NB) distribution, which has been applied to build tools such as edgeR (Robinson *et al*., 2010), DESeq (Anders and Huber, 2010) and baySeq (Hardcastle and Kelly, 2010). Models other than the NB have also been developed. For example, Wu *et al*. (2010) built a tool, referred to as ASC, based on the observation that log fold changes obey the normal distribution. The same assumption has also been adopted by Huang *et al*. (2011). Wang *et al*. (2010) proposed an MA-based method that approximates the distribution of the M value (log fold change) at given A value (log intensity sum) with a normal distribution. Srivastava and Chen (2010) adopted a two-parameter, generalized Poisson model to fit the nucleotide-wise read distribution on each gene independently. Cufflinks (Trapnell *et al*., 2010) adopts the delta method, a commonly used method in microarray expression analysis, to estimate the variance of the log fold change.

When comparing RNA-seq data between two conditions, existing methods calculate a *P*-value for the differential expression of each gene. These *P*-values, however, do not measure how much a gene is expressed in one condition relative to the other. Highly significant *P*-values can result from even miniscule relative differences in gene expression, if the number of sequencing tags available for the comparison is large enough. This type of

*To whom correspondence should be addressed.

result is contrary to what is of fundamental biological interest. In most studies of gene expression, the genes of highest interest are those with large relative differences. The relative difference, or fold change, is a basic and widely used measure for identifying differential gene expression. Unfortunately, the raw fold change is unreliable because it does not take into account the uncertainty of gene expression measures under the two conditions being compared. In particular, the fold changes of genes with low read counts are less reliable than those of genes with high read counts. In other words, fold changes based on read counts are not comparable for genes with different expression levels or genes of different lengths. In microarray analysis, a well-known variance stabilization method is fold change with offset (or start-log) (Durbin *et al*., 2002; Ritchie *et al*., 2007; Rocke and Durbin, 2003), which adds an offset before calculating fold changes. Researchers have also attempted to combine the fold change and *P*-value to provide more meaningful results by setting cutoffs for both the fold change and *P*-value as in the well-known volcano plot (Cui and Churchill, 2003). However, such an *ad hoc* technique does not result in meaningful gene rankings.

Although sequencing costs are rapidly decreasing, experimentalists are reluctant to sequence replicate RNA samples without pilot data that demonstrate the utility of the study. By comparing single RNA-seq treatment samples from control conditions, experimentalists can obtain valuable information that allows them to adjust the experimental plan before scaling up to include multiple replicates. That may partially explain why almost 70% of all the human RNA-seq samples in the gene expression omnibus (Barrett *et al*., 2011) do not have biological replicates and more un-replicated RNA-seq data were published than replicated RNA-seq data in 2011 (Supplementary Fig. S1). Despite the importance of single replicate studies, there is currently no satisfactory method for detecting differentially expressed genes when only a single biological replicate is available.

In this article, we describe a technique for estimating fold change that takes into account the uncertainty of gene expression measurement by RNA-seq. We argue that this new measure of fold change is more informative for the biology of a perturbed system than either *P*-values or raw fold change especially for single biological replicate experiments. From a Bayesian perspective, our representation of fold change is derived from the posterior distribution of the raw fold change. This representation, denoted as GFOLD, balances the estimated degree of change with the significance of this change. GFOLD is more reliable than raw fold change for estimating the relative difference of gene expression and facilitates the comparison of genes with different expression levels or of different lengths.

To validate its effectiveness, we applied GFOLD to several datasets with biological replicates and compared it with edgeR, DESeq, DEGseq, Poisson, Cufflinks and fold change with offset. By comparing the results of different methods when biological replicates are not available with the results when biological replicates are available, we were be able to estimate the performance of different methods using information from single replicates. We also explored the biological significance of gene rankings produced using different methods. Comparisons show that GFOLD outperforms all other methods in most cases when there is only a single replicate. We built a hierarchical model for cases in which biological replicates are available. In such cases, GFOLD provides comparable results to existing methods.

2 METHODS

We describe the single-replicate model in the main text and leave the multiple-replicate model in the supplementary. Because the technical variance of RNA-seq is negligible (Bullard *et al*., 2010), the read count of a gene can be effectively modeled by the Poisson distribution (Jiang and Wong, 2009). Specifically, the probability of observing k short reads associated with a gene is as follows:

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}, \lambda = n \times l \times x \quad (1)$$

where x is the expression level of this gene [(e.g. in RPKM (Reads Per Kilo bases per Million reads) Mortazavi *et al*., 2008)], n is a normalization constant reflecting the sequencing depth and l is the gene length. The method proposed by Anders and Huber (Anders and Huber, 2010) was used to calculate n. If this method fails, we simply treat n as the library size (sequencing depth).

The observed read count k is based on the expression level x, which is actually what we want to measure. From a Bayesian point of view, λ and therefore x can be treated as random variables, and the posterior distribution of λ is defined by

$$\text{Post}(\lambda) \propto \frac{\lambda^k e^{-\lambda}}{k!} \quad (2)$$

which is a gamma distribution with shape $k+1$ and scale 1. Here, the uniform distribution is used as the prior for λ.

For a gene, given the observed read counts under two conditions, the posterior distribution of expression levels x_1 and x_2 under the two conditions can be calculated as above. Furthermore, the posterior distribution of log2 fold change $\log_2(x_2/x_1)$ can also be calculated. Note that the calculation involves l, which can be avoided by calculating the posterior distribution of $y_i = l \times x_i, i \in \{1,2\}$ and $z = \log_2(y_2/y_1)$ instead. It is obvious that $\log_2(y_2/y_1)$ and $\log_2(x_2/x_1)$ have the same distribution.

To utilize the variance information of the posterior distribution of fold change, we define the generalized fold change of a gene as follows:

$$\text{GFOLD}(c) = \begin{cases} \max(t,0) | P_Z(z \le t) = c, & \text{if } mean(Z) \ge 0 \\ \min(t,0) | P_Z(z \ge t) = c, & \text{if } mean(Z) < 0 \end{cases} \quad (3)$$

where P_Z is the posterior distribution of $z = \log_2(y_2/y_1)$ and c is a parameter with default parameter 0.01. If the gene is up-regulated, i.e. $mean(Z) \ge 0$, the probability of the log2 fold change (2nd/1st) being larger than t is $1-c$. In this case, t is less than $mean(Z)$ with default value of c. GFOLD(c) takes $max(t,0)$ to truncate negative t to zero. If GFOLD(c) is 0, then the expression level of this gene does not show significant change. The case for $mean(Z) < 0$ is symmetric to the case of $mean(Z) \ge 0$. Genes can be ranked by their GFOLD values in descending order such that top ranked ones are differentially up-regulated and bottom ranked ones are differentially down-regulated.

Figure 1 shows examples of the posterior distributions of log2 fold change and the calculated GFOLD values for three up-regulated genes. The figure also compared the gene rankings based on the naive read count fold change, GFOLD value and *P*-value for the three genes. According to the *P*-value we would consider the black gene to be the most significant one, followed by green then red. The ranking of black above green is problematic because the true fold change of the green gene is very likely to be much greater than that of the black. If we used the naive fold change to rank genes, green would be the most significant one, followed by red then black. Although the red gene has a greater naive fold change than the black gene, the read counts for the red gene are low and the estimate of fold change is unreliable. GFOLD strikes a balance between fold change and *P*-value, ranking green first, followed by black and

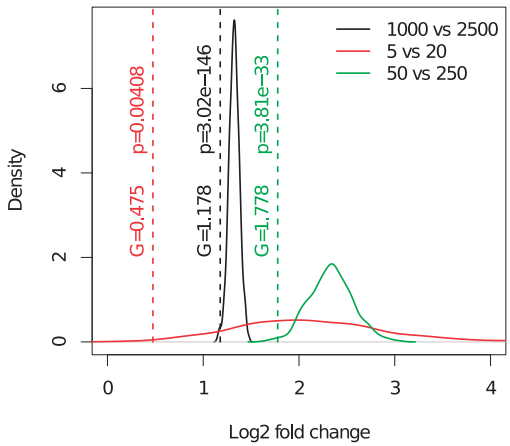

Fig. 1. Rankings of example genes by GFOLD, fold change and *P*-value. The figure illustrates the idea of GFOLD by comparing gene rankings defined by GFOLD (0.01), fold change and *P*-value on three example genes. The read counts of the black, red and green genes are (1000, 2500), (5, 20) and (50, 250) under two biological conditions with the same sequencing depth, respectively. The three curves are the corresponding posterior distributions of log fold change. 'G' stands for GFOLD value and 'p' stands for *P*-value calculated using Poisson test

then red. GFOLD not only measures the fold change but also captures the small variance of the posterior distribution of log2 fold change for the highly expressed genes.

If the fold change based on the read count is fixed and $c < 0.5$, then GFOLD(c) penalizes genes with low expression levels because the posterior distribution of the fold change from such genes has a large variance. This trend is clearly demonstrated in Figure 1. Setting c to a smaller value shifts the preference from genes with higher fold changes at lower expression values to genes with lower fold changes at higher expression levels. By default, we set $c = 0.01$, which means that in 99% of cases, the fold change of a gene is above the absolute GFOLD(0.01) value for this gene.

To better understand the properties of the posterior distribution of fold change and GFOLD, we compared several posterior distributions of fold change with normal distributions in Supplementary Figure S2. When read counts are high, the posterior distributions of fold change are similar to normal distributions; but the similarity disappears when read counts are low.

It might be possible to describe the posterior distribution of log2 fold change with a closed-form formula based on the distribution of the ratio of two gamma variables (Kwan and Leung, 2005). However, the close form calculation is too time consuming to be practical because it involves the inverse of the distribution function. Therefore, in this work, the calculation is done by sampling. Specifically, we first sample $y_i = l \times x_i, i \in \{1, 2\}$ according to Equation (2), then estimate the posterior distribution of $z = \log(y_2/y_1)$ based on sampled values of $y_i, i \in \{1, 2\}$, and last calculate GFOLD(c) according to Equation (3). The whole process is very efficient. For example, to calculate GFOLD values for all the genes given a pair of samples used in this work, it costs less than 30 s on a typical desktop PC.

3 APPLICATIONS

3.1 Datasets

We assessed the effectiveness of our approach on the following five publicly available RNA-seq and GRO-seq datasets that contain biological replicates. For each dataset, we merged technical replicates to form a single biological replicate. The five datasets are as follows: *human brain dataset* downloaded from DDBJ (Kaminuma *et al.*, 2011) with accession number SRA009447, *mouse brain dataset* (Polymenidou *et al.*, 2011) downloaded from DDBJ with accession number SRA030347, *Type I latency B-cell line dataset* (Xu *et al.*, 2010) downloaded from DDBJ with accession number SRP001880, *ENCODE dataset* (Birney *et al.*, 2007) downloaded from UCSC Genome Browser (Kent *et al.*, 2002) and *estrogen response dataset* containing GRO-seq data (Hah *et al.*, 2011) and ERa-binding data (Welboren *et al.*, 2009). The detailed description of each dataset is in the Supplementary Material.

3.2 Methods compared

In the following comparison, we compared GFOLD with the following methods: edgeR (Robinson *et al.*, 2010), DESeq (Anders and Huber, 2010), Cufflinks (Trapnell *et al.*, 2010), DEGseq (Wang *et al.*, 2010), Poisson test and fold change with offset. Because GFOLD value depends on the cutoff c, we use GFOLDX to denote GFOLD with $c = X/1000$ (i.e. GFOLD10 means cutoff 0.01). The method edgeR has an option of whether to estimate the tag-wise dispersion. We denote the version of edgeR that estimates tag-wise dispersion as edgeRT. For DEGseq, there are two MA-based methods: MA-plot-based method with Random Sampling model (MARS) and MA-plot-based method with Technical Replicates (MATR). MARS deals with replicates by summing up read counts for every gene and then treating the data as if there were no replicates, whereas MATR estimates the variance based on replicates. When replicates were available we found the MATR results to be less accurate than those produced by MARS (results now shown); therefore, we used MARS method for DEGseq. Because Poisson test (Poisson) is the simplest method considered for detecting differentially expressed genes, it was included in the comparison. For fold change with offset, because the optimal offset is unknown, we tried different offsets: 1, 5, 10, 20, 30, 40, 50 and 60. Fold change with offset X is denoted as FCX.

3.3 MA plot comparisons

MA plot is a convenient way to display differentially expressed genes. We applied MA plot to compare the results from nine methods (GFOLD5, GFOLD10, FC20, Cufflinks edgeRT, DESeq, edgeR, DEGseq and Poisson) on a pair of samples (one control and one TDP-43 depleted sample) selected from the mouse brain dataset.

Supplementary Figure S3 showed the distribution of the variance and signal strength for the top 1000 differentially expressed genes identified by different methods. DEGseq and edgeR show similar results as Poisson. They are biased towards genes with larger read counts. Such a bias would cause less biological meaningful results as shown in the following sections. GFOLD5, GFOLD10, DESeq and FC20 show less bias.

Although all of the methods except for GFOLD and FC20 adopt *P*-values for testing significance, the *P*-values of different methods are not comparable (as shown in Supplementary Fig. S4). DESeq tends to overestimate the variance and, as a result, identifies fewer differentially expressed

genes using the same *P*-value cutoff. The high similarity between Poisson, edgeR and DEGseq indicates that the later two methods essentially calculate the Poisson variance introduced by random sampling in the sequencing step and thus omit biological variance. Therefore, none of the nine methods properly estimate the biological variance when no biological replicate is available.

3.4 Ranking comparisons

Although we do not know the true levels of gene expression under different conditions, in theory, estimates of the gene expression levels improve with the increase of replicate number. We therefore can evaluate the performance of a method on single replicate against a proxy for the truth in which multiple biological replicates are used. As the different statistics results from different algorithms are not directly comparable, we compared the rankings of the gene differential expression statistics to evaluate the performance of the different approaches while avoiding setting arbitrary cutoffs for different methods.

Specifically, given two top N up-regulated (or N down-regulated) gene rankings, we took the top $k < \mathrm{N}$ genes in each ranking and then calculated the number of genes in common (denoted as k) in the two top gene sets. Then, $\hat{k}/k$, ranging from 0 to 1, measured the similarity of the two top k genes. After all ks were considered, we were able to draw a curve with k/N as the x-axis and $\hat{k}/k$ as the y-axis. We further defined rank area as the area below the curve, similar to the calculation of the area based on an ROC curve. If the rank curve reached the top left corner of the plot, i.e. the rank area was 1, then the two gene rankings were identical. The rank area is a direct measure of the similarity of the two rankings. We set $N = 1000$ in the comparisons. Figure 2 illustrates an example that compares the results produced by DESeq on all replicates (as benchmark) with the results by all methods on only a pair of single replicates on Type I B-cell dataset. GFOLD in general gives the better overlap with the benchmark, with the default GFOLD10 giving the best results. Compared with FC, GFOLD is less sensitive to the parameters as shown in Supplementary Figure S5.

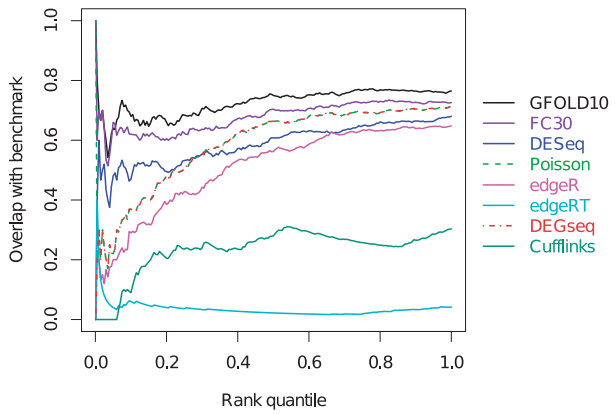

Fig. 2. An example of the rank area comparison. The up-regulated gene rankings of different methods on a sample pair are compared with the gene ranking by DESeq given all replicates on the Type I B-cell dataset. A curve that gives larger area under the curve is considered better

Of all the methods we compared, six (Cufflinks, GFOLD, DEGseq, edgeR, DESeq and edgeRT) accepted replicates. Theoretically, results produced by each of them on all replicates could be used as benchmark, but we did not know which derived benchmark was the most reliable one. Therefore, we compared the results of each method on every possible pair of single replicates with all derived benchmarks from six methods separately. The comprehensive comparison on human brain datasets is given in the heatmaps illustrated in Supplementary Figure S6. Except for using DEGseq as benchmark, GFOLD in general performed better than any given method even when the benchmark was produced by that method. When the benchmark by DEGseq was used, DEGseq, edgeR and Poisson gave better results than GFOLD. We looked into details of the results and found that the MARS method adopted by DEGseq summed up read counts from replicates and behaved similar to Poisson and edgeR, as shown in Supplementary Figure S3(G–I). Therefore, edgeR, DEGseq and Poisson performed better when the benchmark by DEGseq was used. Supplementary Figure S6 also shows that the performance of FC10 was comparable to that of GFOLD10. To better understand the difference between GFOLD and fold change with offset, we further evaluated the performance of different methods on other datasets. For each method on each dataset, Table 1 shows the average rank area which was calculated by averaging over all possible sample pairs, all benchmarks and the up-/down-regulated gene lists. In the table, methods are ranked by their average rank areas over all datasets. For each dataset, fold change with the optimal offset was comparable with GFOLD10. However, the optimal offset for fold change varied for different datasets. For example, FC10 performed better than other offsets in the human brain and the ENCODE dataset but not on the other three datasets. Furthermore, the selection of the optimal offset for fold change seems uncorrelated with the sequencing depth. To the contrary, GFOLD with parameter 0.005 or the default parameter 0.01 performs well on all five datasets.

3.5 In-group and out-group comparisons

Given a dataset with biological replicates under two conditions, an in-group comparison consists of all possible comparisons between two biological replicate samples. An out-group comparison consists of all possible comparisons between two samples from different conditions. A commonly used strategy to assess the specificity of a method is to do an in-group and out-group comparison and to identify differentially expressed genes using the same cutoff. If a method reports fewer differentially expressed genes in the in-group comparison but reports more differentially expressed genes in the out-group comparison, then such a method is considered to achieve better specificity.

In previous studies, the in-group and out-group comparisons were conducted by setting a default cutoff for the *P*-value. However, as we stated above, *P*-values used by different methods are not comparable. Furthermore, GFOLD and fold change with offset do not rank genes by *P*-values. To make a fair comparison of different methods using the in- and out-group comparisons, we adopted a strategy similar to the calculation of FDR mentioned above. For a method generating a *P*-value, we first applied this method to all possible in- and out-group

Table 1. Average rank area of different methods on all five datasets

Methods (Depth M)	HB 7.8–26.4	EN 11.1–18.4	MB 4.7–9.7	TIB 8.4–28.9	ER 4.8–10.9	Ave
GFOLD5	0.584	0.712	0.513	0.470	0.692	0.594
GFOLD10	0.585	0.710	0.515	0.470	0.687	0.593
FC20	0.567	0.707	0.505	0.442	0.693	0.583
GFOLD50	0.584	0.703	0.508	0.449	0.661	0.581
FC30	0.552	0.695	0.510	0.453	0.694	0.581
FC40	0.538	0.683	0.509	0.457	0.688	0.575
FC50	0.527	0.672	0.506	0.456	0.681	0.568
FC10	0.583	0.714	0.478	0.398	0.663	0.567
GFOLD100	0.580	0.696	0.493	0.419	0.638	0.565
FC60	0.516	0.662	0.503	0.454	0.673	0.561
DESeq	0.559	0.645	0.496	0.401	0.627	0.546
FC5	0.579	0.703	0.428	0.321	0.599	0.526
Poisson	0.354	0.453	0.419	0.369	0.539	0.427
edgeR	0.348	0.440	0.426	0.376	0.544	0.427
DEGseq	0.345	0.410	0.420	0.370	0.537	0.416
Cufflinks	0.468	0.523	0.398	0.251	N/A	0.410
FC1	0.505	0.598	0.238	0.098	0.366	0.361
edgeRT	0.235	0.393	0.166	0.089	0.474	0.271

For each dataset, the abbreviation and the range of the number of mappable reads (in million reads) for each sample are as follows. HB: human brain (7.8–26.4); EN: ENCODE (11.1–18.4); MB: mouse brain (4.7–9.7); TIB: Type I B-cell (8.4–28.9); ER: estrogen response (4.8–10.9); Ave: average.

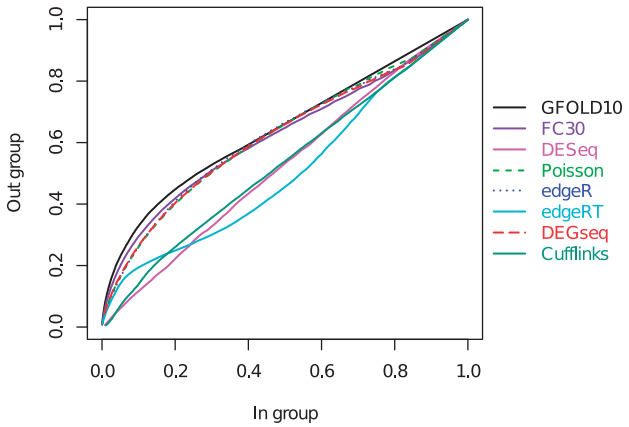

Fig. 3. The in-out-group comparison on the human brain dataset. The *x*-axis is the percentage of genes above (for GFOLD and fold change with offset) or below (for other methods) a certain cutoff in the in-group comparison. The *y*-axis is the percentage of genes above/blow a certain cutoff in the out-group comparison. A curve that is closer to the top left corner of the plot is considered to achieve better specificity

comparisons and determined all of the *P*-values assigned to every gene. Then, we drew a curve with the *x*-axis representing the percentage of genes with *P*-values below a certain cutoff in the in-group comparison and the *y*-axis representing the percentage of genes with *P*-values below the same cutoff in the out-group comparison. This approach was similar for GFOLD and fold change with offset, except that we calculated the percentage of genes with absolute fold changes or GFOLD values above certain cutoffs.

Figure 3 and Supplementary Figure S7 represent the results for the human brain dataset. The figures show that GFOLD achieved the best specificity when a large GFOLD value cutoff was used, which is a desired property because we are more interested in these genes. Supplementary Figure S7 also shows that GFOLD is not sensitive to the parameter. DEGseq, edgeR and Poisson produced similar curves, which is consistent with the previous analysis that DEGseq and edgeR give similar results with those by Poisson. Although fold change with offset performed better than GFOLD on the Type I B-cell dataset and edgeRT performed slightly better than GFOLD on the estrogen response dataset, GFOLD generally performed well and is not sensitive to the cutoff (as shown in Supplementary Figure S8).

In general, the difference between two biological replicates is expected to be smaller than that between two samples from different conditions. If the variance is not overestimated, as in the Poisson model, under the same *P*-value cutoff, we expected to observe more genes with more significant *P*-values when comparing two samples from different conditions. The fact that the curves for edgeR, DESeq and Cufflinks are close to the reverse diagonal line indicates that under the same *P*-value cutoff the three methods give similar number of differential genes for both in- and out-group comparisons, demonstrating that edgeRT, DESeq and Cufflinks overestimated the variance for out-group comparison. Therefore, for those three methods, more stringent cutoffs could be selected to identify more differently expressed genes. In practical data analysis, the gene ranking is often of greater importance that the *P*-values themselves; in many downstream data analyses, such as gene set enrichment analysis, a ranking of genes is the only required input (Subramanian *et al.*, 2005).

3.6 The biological significance of gene rankings on ENCODE datasets

Although previous comparisons were concerned with consistency within datasets these comparisons do not address the main reason for using GFOLD that it provides a useful measure of biologically relevant changes. This section focuses on the biological significance of gene rankings given by different methods using the functional annotation service provided by DAVID (Dennis *et al.*, 2003).

We first checked the functional annotations produced by DAVID given the top 1000 genes up-regulated in K562 using all biological replicates in both K562 and GM12878 by six different methods. Generally, the top-ranked functional annotations are very similar for results from GFOLD, DESeq, edgeR and edgeRT both in ranking and significance. Except for DEGseq, the first functional annotation clusters by the other methods contain the same four Gene Ontology (GO) terms: angiogenesis, vasculature development, blood vessel development and blood vessel morphogenesis. For GFOLD, DESeq, edgeR and edgeRT, the enrichment scores of this cluster are all over 5 and the Benjamini *P*-values for every GO term are all below 0.01. High enrichment of the angiogenesis related cluster is an expected result, because increased angiogenesis is one of the characteristics of chronic myelogenous leukemia cell line like K562 (Vidović *et al.*, 2009; Zhelyazkova *et al.*, 2008).

We then evaluated the performance of all methods on pairs of single replicates. The top 1000 up-regulated genes produced by each method were used to do functional annotation analysis as above. We checked whether the angiogenesis-related set could be identified with high significance (enrichment scores >5 and Benjamini value <0.01) by DAVID given the results from each method. The DAVID results, summarized in Supplementary Table S1, showed that only results produced by GFOLD5, GFOLD10, GFOLD50, GFOLD100 and FC5 could pass the test on both pairs of single replicates simultaneously. Furthermore, in most of the cases, the angiogenesis-related set is ranked top for GFOLD and fold change with offset. Therefore, when only a single replicate is available, GFOLD and fold change with offset could give more biological meaningful gene rankings, and results by GFOLD were more stable and less sensitive to the parameter than that by fold change with offset.

We further redid the GO analysis using GOseq (Young *et al*., 2010) to avoid possible bias due to over-detection of differential expression for long and highly expressed transcripts. Similar to the results of DAVID, the GOseq results, summarized in Supplementary Tables S2–S4, show that blood vessel development and blood vessel morphogenesis are ranked top given the gene lists produced by GFOLD, DESeq, edgeR, edgeRT and Cufflinks when all biological replicates are used. When single replicate is used, GFOLD, FC and Cufflinks give these two GO terms much higher and more consistent ranks than the other methods do, and GFOLD is less sensitive to the parameter compared with FC.

3.7 The biological significance of gene rankings on estrogen response dataset

We further discuss the biological significance of gene rankings based on the estrogen response dataset. Using the breast cancer cell line MCF7, Hah *et al*. (2011) performed a GRO-seq experiment that measures the rate of transcription, different from the cellular concentration of mRNA that is measured by RNA-seq. Comparing transcription rates before and after estrogen stimulation, they found that genes with estrogen induced transcription rates tended to have estrogen receptor (ER) binding sites near their transcription start sites (TSSs). This is consistent with the biological mechanism in which ER binding at enhancers interacts with the transcriptional machinery at the TSS to induce transcription. Gene rankings that have the largest proportion of ER binding near the top of the list could be said to be of the greatest biological relevance. In Supplementary Figure S9 (A–C), we compare, as a function of ranking, the proportion of genes having at least one ER binding site within 10 kb of the TSS. Here, for a better visualization, we used the default cutoff for GFOLD and selected 30 as the optimal offset for fold change based on Table 1. When two biological replicates are used [Supplementary Fig. S9 (A)], all methods achieve a similar performance except for DEGseq. DESeq and edgeRT exhibit a clear drop when the top 100–200 genes are considered. When a pair of single replicates is used [Supplementary Fig. S9 (B and C)], GFOLD10, FC30 and DESeq have similar behavior, clearly outperforming the other methods.

4 DISCUSSION

An experiment with multiple treatment and control replicates should be considered better designed and more informative than one that has single treatment and control samples. However, it is common for experimental groups to have biological material that is sufficient for only single replicate RNA-seq experiments. The data from such experiments provide imperfect information which is nevertheless far better than no information at all. In many experimental settings when it is not clear which downstream experiment is better, single replicate experiments are efficient for collecting useful knowledge that allows informed decisions to be made. In this study, we have shown that GFOLD provides a more consistent and more biologically meaningful approach to ranking differentially expressed genes than other commonly used methods for RNA-seq data without biological replicates, Although raw fold change does not take the variance of gene expression into account, and many *P*-value estimation methods rely too much on variance estimates, GFOLD seeks a balance between estimates of expression change and variability. The GFOLD value for each gene can be considered as a robust fold change, which measures primarily the relative change of the expression level instead of the significance (i.e. *P*-value) of differential expression. The evaluation results conducted on real datasets showed that in the analysis of a single pair of replicates, the gene ranking by GFOLD is better than other methods in terms of consistency with rankings based on multiple replicate data. The results from the ENCODE and GRO-seq dataset suggest that when using single replicate data the rankings produced by GFOLD are more biologically meaningful than the gene rankings by other methods. We stress that while useful information can be obtained from single replicate experiments, an analysis of variation between samples is necessary to draw sound conclusions, especially on the individual gene level.

In comparison with other methods, besides fold change with offset, GFOLD is simple and intuitive. If there is no biological replicate, GFOLD is based on only one assumption that the read count of a gene follows a Poisson distribution. On the contrary, *P*-value is likely to be inaccurate without reliable measurements and modeling of gene expression variation because it is in the tails of the null distribution. Even if the biological variance has been captured accurately, *P*-values from other methods tell only of whether a gene is differentially expressed and not directly of the relative difference in expression. In comparison with GFOLD, fold change with offset is even simpler and could give comparable results if the optimal offset is selected; however, it is hard to select the optimal offset. Furthermore, the optimal offset seems uncorrelated with the sequencing depth. In contrast, GFOLD with the default cutoff performed well on all five datasets we have tested.

GFOLD is generalizable to isoform-level analysis of differential expression. The main challenge is that a system of linear equations has to be solved before calculating the posterior distribution of log fold change (Jiang and Wong, 2009), which is out of scope of this study. In general, we believe that the ranking of gene expression changes by GFOLD, using a reliable measure of fold change, is more biologically meaningful than ranking by *P*-value. This concept can be broadly applied, beyond RNA-seq or GRO-seq, to other types of genomic data, including ChIP-seq.

Funding: The National Basic Research Program of China (973 Program; 2010CB944904 and 2011CB965104), the National Natural Science Foundation of China (31028011 and 31071114), the Shanghai Rising-Star Program (10QA1407300), the New Century Excellent Talents in the University of China (NCET-11-0389), Shanghai Key Laboratory of Intelligent Information Processing of China (20102662), the Excellent Young Teachers Program of Tongji University (2010KJ041) and the Innovative Research Team Program Ministry of Education of China (IRT1168).

Conflict of Interest: none declared.

REFERENCES

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

Barrett,T. *et al.* (2011) NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.*, **39** (Database issue), D1005–1010.

Birney,E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Bullard,J.H. *et al.* (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.

Cloonan,N. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, **5**, 613–619.

Cui,X. and Churchill,G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.*, **4**, 210.

Dennis,G., Jr *et al.* (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.

Durbin,B.P. *et al.* (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18** (**Suppl. 1**), S105–S110.

Haas,B.J. and Zody,M.C. (2010) Advancing RNA-Seq analysis. *Nat. Biotech.*, **28**, 421–423.

Hah,N. *et al.* (2011) A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell*, **145**, 622–634.

Hardcastle,T. and Kelly,K. (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.

Huang,W. *et al.* (2011) Efficiently identifying genome-wide changes with next-generation sequencing data. *Nucleic Acids Res.*, **39**, pe130.

Jiang,H. and Wong,W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026–1032.

Kaminuma,E. *et al.* (2011) DDBJ progress report. *Nucleic Acids Res.*, **39** (Database issue), D22–D27.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Kwan,R. and Leung,C. (2005) Gamma variate ratio distribution with application to CDMA performance analysis. *Advances in Wired and Wireless Communication, 2005 IEEE/Sarnoff Symposium on*, 188–191.

Lister,R. *et al.* (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.

Marioni,J.C. *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.

Morin,R. *et al.* (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques*, **45**, 81–94.

Morozova,O. *et al.* (2009) Applications of new sequencing technologies for transcriptome analysis. *Annu. Rev. Genom. Hum. Genet.*, **10**, 135–151.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Nagalakshmi,U. *et al.* (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.

Polymenidou,M. *et al.* (2011) Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nat. Neurosci.*, **14**, 459–468.

Ritchie,M.E. *et al.* (2007) A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, **23**, 2700–2707.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Rocke,D.M. and Durbin,B. (2003) Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics*, **19**, 966–972.

Srivastava,S. and Chen,L. (2010) A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res.*, **38**, e170.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U S A*, **102**, 15545–15550.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotech.*, **28**, 511–515.

Vidović,A. *et al.* (2009) Prognostic significance of cellular vascular endothelial growth factor (VEGF) expression in the course of chronic myeloid leukaemia. *Srpski Arhiv Za Celokupno Lekarstvo*, **137**, 379–383.

Wall,P.K. *et al.* (2009) Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genom.*, **10**, 347.

Wang,L. *et al.* (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**, 136–138.

Wang,Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.

Welboren,W.-J. *et al.* (2009) ChIP-Seq of ER[alpha] and RNA polymerase II defines genes differentially responding to ligands. *EMBO J.*, **28**, 1418–1428.

Wilhelm,B.T. *et al.* (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.

Wu,Z. *et al.* (2010) Empirical bayes analysis of sequencing-based transcriptional profiling without replicates. *BMC Bioinformatics*, **11**, 564.

Xu,G. *et al.* (2010) Transcriptome and targetome analysis in MIR155 expressing cells using RNA-seq. *RNA*, **16**, 1610–1622.

Young,M.D. *et al.* (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.*, **11**, R14.

Zhelyazkova,A.G. *et al.* (2008) Prognostic significance of hepatocyte growth factor and microvessel bone marrow density in patients with chronic myeloid leukaemia. *Scand. J. Clin. Lab. Invest.*, **68**, 492–500.