RESEARCH

Bootstrap-based differential gene expression analysis for RNA-Seq data without replicates

Sahar Al Seesi[1*†], Yvette Temate Tiagueu[2†], Alexander Zelikovsky[2] and Ion I. Măndoiu[1]

*Correspondence: sahar@engr.uconn.edu
[1]Computer Science & Engineering Department, University of Connecticut, 06269 Storrs, CT, USA
Full list of author information is available at the end of the article
†Equal contributor

Abstract

A major application of RNA-Seq is to perform differential gene expression analysis. Many tools exist to analyze differentially expressed genes in the presence of biological replicates. Frequently, however, RNA-Seq experiments have no or very few biological replicates and development of methods for detecting differentially expressed genes in these scenarios is still an active research area.

In this paper we introduce a novel method, called IsoDE, for differential gene expression analysis based on bootstrapping. We compared IsoDE against four existing methods: Fisher's exact test, GFOLD, edgeR and Cuffdiff on RNA-Seq datasets generated using three different sequencing technologies, both with and without replicates. Experiments on MAQC RNA-Seq datasets without replicates show that IsoDE has consistently high accuracy as defined by the qPCR ground truth, frequently higher than that of the compared methods, particularly for low coverage data and at lower fold change thresholds. In experiments on RNA-Seq datasets with up to 7 replicates, IsoDE has also achieved high accuracy. Furthermore, unlike GFOLD and edgeR, IsoDE accuracy varies smoothly with the number of replicates, and is relatively uniform across the entire range of gene expression levels.

The proposed non-parametric method based on bootstrapping has practical running time, and achieves robust performance over a broad range of technologies, number of replicates, sequencing depths, and minimum fold change thresholds.

Keywords: differential gene expression; bootstrapping; RNA-Seq

Introduction

RNA-Seq has become the new standard for the analysis of differential gene expression [1] [2] [3] due to its wider dynamic range and smaller technical variance [4] compared to 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 read mapping uncertainty or capture, fragmentation, and amplification variability in library preparation and sequencing. Therefore, the need for using statistical methods arises. Traditionally, statistical methods rely on the use of replicates to estimate biological and technical variability in the data. Popular methods for analyzing RNA-Seq data with replicates include edgeR [5], DESeq [6], Cuffdiff [7], and the recent NPEBSeq [8].

Unfortunately, due to the still high cost of sequencing, many RNA-Seq studies have no or very few replicates [9]. Methods for performing differential gene expression analysis of RNA-Seq datasets without replicates include variants of Fisher's

exact test [4]. Recently, Feng et al. introduced GFOLD [10], a non-parametric empirical Bayesian-based approach, and showed that it outperforms methods designed to work with replicates when used for single replicate datasets.

In this work, we present a novel method for differential gene expression analysis for RNA-Seq data, called IsoDE. Our method uses the traditional bootstrapping approach [11] to resample RNA-Seq reads, in conjunction with the accurate Expectation-Maximization IsoEM algorithm [12] to estimate gene expression levels from the samples. Experimental results on RNA-Seq datasets generated using three different technologies (Illumina, ION Torrent, and 454) from two well-characterized MAQC [13] samples show that IsoDE has consistently high accuracy, comparable or better than that of Fisher's exact test, GFOLD, Cuffdiff, and edgeR (we did not compare directly with NPEBSeq since installation was not successful). Notably, and unlike other methods, IsoDE maintains high accuracy (sensitivity and PPV around 80%) on low coverage RNA-Seq datasets and at lower fold change thresholds.

Recent studies such as Rapaport et al. [14] have reiterated the fact that increasing the number of replicate samples significantly improves detection power over increased sequencing depth. We explored the effect of the number of replicates on prediction accuracy using a RNA-Seq dataset [15] with 7 replicates for each of two conditions (control and E2-treated MCF-7 cells). Although all methods generally benefit from the use of additional replicates, GFOLD and edgeR show a marked discontinuity when transitioning from 1 to 2 replicates. In contrast, IsoDE accuracy varies smoothly with changes in the number of replicates.

Methods

Bootstrap sample generation

As most differential expression analysis packages, IsoDE starts with a set A of RNA-Seq read alignments for each condition. Bootstrapping can be used in conjunction with any method for estimating individual gene expression levels from aligned RNA-Seq reads, estimation typically expressed in *fragment per kilobase of gene length per million reads* (FPKM). In IsoDE, we use the IsoEM algorithm [16], an expectation-maximization (EM) algorithm that takes into account gene isoforms in the inference process to ensure accurate length normalization. Unlike some of the existing estimation methods, IsoEM uses non-uniquely mapped reads, relying on the distribution of insert sizes and base quality scores (as well as strand and read pairing information if available) to probabilistically infer their origin. Previous experiments have shown that IsoEM yields highly accurate FPKM estimates with lower runtime compared to other commonly used inference algorithms [17].

The first step of IsoDE is to generate M bootstrap samples by randomly resampling with replacement from the reads represented in A. When a read is selected during resampling, all its alignments from A are included in the bootstrap sample. The number of resampled reads in each bootstrap sample equals the total number of reads in the original sample. However, the total number of alignments may differ between bootstrap samples, depending on the number of alignments of selected reads and the number of times each read is selected. The IsoEM algorithm is then run on each bootstrap sample, resulting in M FPKM estimates for each gene. The bootstrap sample generation algorithm is summarized below:

1 Sort the alignment file A by read ID
2 Compute the number N of reads and generate a list $\mathcal{L}$ containing read IDs in the alignment file A
3 For $i = 1, \ldots, M$ do:
   (a) Randomly select with replacement N read IDs from $\mathcal{L}$, sort selected read IDs, and extract in A_i all their alignments with one linear pass over A (if a read is selected m times, its alignments are repeated m times in A_i)
   (b) Run IsoEM on A_i to get the i^{th} FPKM estimate for each gene

Bootstrap-based testing of differential expression

To test for differential expression, IsoDE takes as input two folders which contain FPKM estimates from bootstrap samples generated for the two conditions to be compared. In case of replicates, a list of bootstrap folders can be provided for each condition (one folder per replicate, normally with an equal number of bootstrap samples) – IsoDE will automatically merge the folders for the replicates to get a combined folder per condition, then perform the analysis as in the case without replicates.

In the following we assume that a total of M bootstrap samples is generated for each of the compared conditions. We experimented with two approaches for pairing the FPKMs estimated from the two sets of bootstrap samples. In the "matching" approach, a random one-to-one mapping is created between the M estimates of first condition and the M estimates of the second condition. This results in M pairs of FPKM estimates. In the "all" approach, M^2 pairs of FPKM estimates are generated by pairing each FPKM estimate for first condition with each FPKM estimate for second condition. When pairing FPKM estimate a_i for the first condition with FPKM estimate b_j for the second condition, we use a_i/b_j as an estimate for the fold change in the gene expression level between the two conditions. The "matching" approach thus results in $N = M$ fold change estimates, while the "all" approach results in $N = M^2$ fold change estimates.

The IsoDE test for differential expression requires two user specified parameters, namely the minimum fold change f and the minimum bootstrap support b. For a given threshold f (typically selected based on biological considerations), we calculate the percentage of fold change estimates that are equal to or higher than f when testing for overexpression, respectively equal to or lower than $1/f$ when testing for underexpression. If this percentage is higher than the minimum bootstrap support b specified by the user then the gene is classified as differentially expressed (DE), otherwise the gene is classified as non-differentially expressed (non-DE). The actual bootstrap support for fold change threshold f, as well as the minimum fold change with bootstrap support of at least b are also included in the IsoDE output to allow the user to easily increase the stringency of the DE test.

As discussed in the results section, varying the bootstrap support threshold b allows users to achieve a smooth tradeoff between sensitivity and specificity for a fixed fold change f (see, e.g., Figure 1). Since different tradeoffs may be desirable in different biological contexts, no threshold b is universally applicable. In our experiments we computed b using a simple binomial model for the null distribution of fold change estimates and a fixed significance level $\alpha = 0.05$. Specifically, we

assume that under the null hypothesis the fold changes obtained from bootstrap estimates are equally likely to be greater or smaller than f. We then compute b as x_{min}/N, where $x_{min} = \min\{x \ : \ P(X \geq x) \leq \alpha\}$ and X is a binomial random variable denoting the number of successes in N independent Bernoulli trials with success probability of 0.5. For convenience, a calculator for computing the bootstrap support needed to achieve a desired significance level given the (possibly different) numbers of bootstrap samples for each condition has been made available online (see Availability).

The number M of bootstrap samples is another parameter that the users of IsoDE must specify. As discussed in the results section, computing the bootstrap support for all genes takes negligible time, and the overall running time of IsoDE is dominated by the time to complete the $2M$ IsoEM runs on bootstrap samples. Hence, the overall runtimes grows linearly with M. Experimental results suggest that the "all" pairing approach produces highly accurate results with relatively small values of M (e.g., $M = 20$), and thus results in practical runtimes, independent of the number of replicates. We also note that for studies involving pairwise DE analysis of more than two conditions, IsoDE only requires M independently generated bootstrap samples per condition. Since the time for computing pairwise bootstrap support values is negligible, the overall running time will grow linearly with the number of conditions.

Compared methods

The four methods that were compared to IsoDE are briefly described below.

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 organized 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 conditions A and B by setting the two values in the first row of the table to the estimated number of reads mapped per kilobase of gene length (calculated from IsoEM estimated FPKM values) in conditions A and B, respectively. The values in the second row of the contingency table depend on the normalization method used. We compared three normalization methods. The first one is total read normalization, where the total number of mapped reads in conditions 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 condition is used. We also test normalization by ERCCs RNA spike-in controls [18]. 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. In our experiments, we used POLR2A as a housekeeping gene.

The calculated p-value, which measures the significance of deviation from the null hypothesis that the gene is not differentially expressed, is computed exactly by using the hypergeometric probability of observed 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 using the Benjamini and Hochberg method [19] with 5% false discovery rate (FDR).

GFOLD

GFOLD [10] is a generalized fold change algorithm which produces 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. The authors show that GFOLD outperforms other commonly used methods when used for single replicate datasets. We used GFOLD v1.0.7 with default parameters and fold change significance cutoff of 0.05.

Cuffdiff

Cuffdiff [7] uses a beta negative binomial distribution model to test the significance of change between samples. The model accounts for both uncertainty resulting from read mapping ambiguity and cross-replicate variability. Cuffdiff reports fold change in gene expression level along with statistical significance. In our comparison, we used Cuffdiff v2.0.1 with default parameters.

edgeR

edgeR [5] is a statistical method for differential gene expression analysis which is based on the negative binomial distribution. Although edgeR is primarily designed to work with replicates it can also be run on datasets with no replicates. We used edgeR on counts of uniquely mapped reads, as suggested in [15]. We followed the steps provided in the edgeR manual for RNA-Seq data. calcNormFactors(), estimateCommonDisp(), estimateTagwiseDisp(), and exactTest() were used with default parameter, when processing the MCF-7 replicates. When processing MAQC data and a single replicate of MCF-7 data, estimateTagwiseDisp() was not used, and the dispersion was set to 0 when calling exactTest(). The results where adjusted for multiple testing using the Benjamini and Hochberg method with 5%.

Mapping RNA-Seq reads

MAQC Illumina reads were mapped onto hg19 Ensembl 63 transcript library; all other datasets were mapped onto hg19 Ensembl 64 transcript library. Illumina datasets (MAQC and MCF-7) were mapped using Bowtie v0.12.8 [20]. ION Torrent reads were mapped using TMAP v2.3.2, and 454 reads were mapped using MOSAIK v 2.1.33 [21]. For edgeR, non-unique alignments were filtered out, and read counts per gene were generated using coverageBed (v2.12.0). Read mapping statistics are detailed in Table 5. Number of mapped reads per kilobase of gene length used in Fisher's exact test calculation are based on IsoEM FPKMs.

Ground truth definition

On MAQC dataset the ground truth was defined based on the available qPCR data from [13]. Each TaqMan assay was run in four replicates for each measured gene. POLR2A (ENSEMBL gene 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: $2^{(\text{CT of POLR2A})-(\text{CT of g})}$. We filtered out genes that: (1) were not detected in one 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 unambiguously mapped to Ensemble gene IDs (686 genes). A gene was considered differentially expressed if the fold change between the average normalized TaqMan expression levels bin the two conditions was greater than a set threshold with the p-value for an unpaired two-tailed T-test (adjusted for 5% FDR) of less than 0.05. We ran the experiment for fold change thresholds of 1, 1.5, and 2.

For experiments with replicates we used the RNA-Seq data generated from E2-treated and control MCF-7 cells in [15]. In this experiment, we compared IsoDE with GFOLD and edgeR. The predictions made by each method when using all 7 replicates for each condition were used as its own ground truth to evaluate predictions made using fewer replicates. The ground truth for IsoDE was generated using a total of 70 bootstrap samples per condition.

Evaluation metrics

For each evaluated method, 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, and accuracy, defined as follows:

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

Results and discussion

Datasets

We conducted experiments on publicly available RNA-Seq datasets generated from two commercially available reference RNA samples and a breast cancer cell line.

To compare the accuracy of different methods, we used RNA-Seq data RNA samples that were well-characterized by quantitative real time PCR (qRT-PCR) as part of the MicroArray Quality Control Consortium (MAQC) [13]; 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 as described in the Methods section from TaqMan qRT-PCR measurements collected as part of the MAQC project (GEO accession GPL4097).

We used RNA-Seq data generated for HBRR and UHRR using three different technologies: Illumina, ION-Torrent, and 454. Details about the datasets and their SRA accession numbers (or run IDs for ION Torrent datasets) are available in Table 5 in Additional File 1. The ION Torrent runs for each MAQC sample were combined to generate the results in Table 3, results obtained using pairs of individual ION Torrent runs are reported in Tables 7 and 8 in Additional File 1.

The MCF-7 RNA-Seq data was generated (from the MCF-7 ATCC human breast cancer cell line) by Liu et al. [15] using Illumina single-end sequencing with read length of 50bp. A total of 14 biological replicates were sequenced from two conditions: 7 replicates for the control group and 7 replicates for E2-treated MCF-7 cells. Sequencing each replicate resulted produced between 25 and 65 millions of mapped reads. Details about this dataset and accession numbers are also available in Table 5 in Additional File 1.

Bootstrapping support and pairing strategy effects on IsoDE accuracy and runtime

We evaluated both the "matching" and "all" pairing strategies of IsoDE (referred to as IsoDE-Match and IsoDE-All) for fold change threshold f of 1, 1.5, respectively 2, and bootstrap support threshold b between 40% and 95%. The results of IsoDE-Match with $M = 200$ bootstrap replicates per condition are shown in Figure 1. The results show that, for each tested value of f, varying b results in a smooth tradeoff between sensitivity and PPV, while the F-score changes very little. For the remaining experiments we used a bootstrap support level b computed using a significance level of 0.05 under the binomial null model detailed in the Methods section. Note the value of b selected in this way depends on the number of number N of fold change estimates, which in turn depends on both M and the pairing strategy (N is equal to M for IsoDE-Match, respectively to M^2 for IsoDE-All).

To determine the best pairing strategy, we ran IsoDE-Match and IsoDE-All with number of bootstrap samples M varying between 10 and 200 (results not shown). For the considered measures, IsoDE-All achieved an accuracy very close to that of IsoDE-Match when run with a comparable value of N. For example, as shown in Tables 2-4, IsoDE-All run on $M = 20$ bootstrap samples ($N = 400$) had similar accuracy with the largest number of bootstrap samples we could use with IsoDE-Match ($M = N = 200$).

Since for a fixed N IsoDE-Match requires $2N$ bootstrap samples while IsoDE-All requires only $2\sqrt{N}$ of them, using IsoDE-All is significantly faster in practice. Indeed, most of the IsoDE time is spent generating bootstrap samples and estimating expression levels for each of them using the IsoEM algorithm, with bootstrap support computation typically taking a fraction of a minute. Figure 2 shows the time required to generate $M = 20$, respectively $M = 200$, bootstrap samples for both conditions of several MAQC datasets. All timing experiments were conducted on a Dell PowerEdge R815 server with quad 2.5GHz 16-core AMD Opteron 6380 processors and 256Gb RAM running under Ubuntu 12.04 LTS. IsoEM is run on bootstrap samples sequentially, but for each run its multi-threaded code takes advantage of all available cores (up to 64 in our experimental setup). As expected, the running time scales linearly with the number of bootstrap samples per condition, and thus generating $M = 20$ bootstrap samples per condition is nearly 10 times faster than

generating $M = 200$ of them. Overall, IsoDE-Match with $M = 20$ has reasonable running time, varying between 1 hour for the smallest 454 dataset to 3.5 hours for the Illumina dataset.

Results for DE prediction without replicates

We compared IsoDE against GFOLD, Cuffdiff, edgeR, and different normalization methods for Fisher's exact test; namely total normalization, housekeeping gene (POLR2A) normalization, and normalization using External RNA Controls Consortium (ERCC) RNA spike-in controls [18]. Cuffdiff results were considerably worse on the Illumina MAQC dataset, compared to other methods. Consequently, Cuffdiff was not included in other comparisons. edgeR was also not included in further comparisons due to lack of clear definition of uniquely mapped reads for ION-Torrent and 454 datasets which were mapped using local read alignment tools. ERCC spike-ins were available only for ION Torrent samples; therefore, ERCC normalization for Fisher's exact test was conducted only for ION Torrent datasets.

Table 2 shows the results obtained for the MAQC Illumina dataset using minimum fold change threshold f of 1, 1.5, and 2, respectively. Table 3 shows the results obtained for the combined ION Torrent MAQC dataset for the same values of f. Table 4 shows the results for the First 454 MAQC dataset, while results for the Second 454 dataset are presented in Table 6 in Additional File 1. For each fold change threshold, the best performing method for each statistic is highlighted in bold.

IsoDE has very robust performance, comparable or better than that of the other methods for differential gene expression. Indeed, IsoDE outperforms them in a large number of cases, across datasets and fold change thresholds. Very importantly, unlike GFOLD and Fisher's exact test, IsoDE maintains high accuracy (sensitivity and PPV around 80%) on datasets with small numbers of mapped reads such as the two 454 datasets. This observation is confirmed on results obtained for pairs of individual ION-Torrent runs, presented in Tables 7 and 8 in Additional File 1. This makes IsoDE particularly attractive for such low coverage RNA-Seq datasets.

DE prediction with replicates

We also studied the effect of the number of biological replicates on prediction accuracy using the MCF-7 dataset. We performed DE predictions using an increasing number of replicates. IsoDE was run with a total of 20 bootstrap samples per condition, distributed equally (or as close to equally as possible) among the replicates, as detailed in Table 9. GFOLD and edgeR were evaluated for 1 through 6 replicates using as ground truth the results obtained by running each method on all 7 replicates (see the Methods section). For IsoDE, we also include the results using $M = 20$ bootstrap samples from all 7 replicates as its ground truth is generated using a much larger number of bootstrap samples ($M = 70$). Figure 3 shows the results of the three compared methods for a fold change threshold of 1, results for fold change thresholds 1.5 and 2 are shown in Figures 5 and 6 in Additional File 1.

Since for this experiment the ground truth was defined independently for each method, it is not meaningful to directly compare accuracy metrics of different methods. Instead, we focus on the rate of change in the accuracy of each method as additional replicates are added. Generally, all methods perform better when increasing

the number of replicates. However, the accuracy of IsoDE varies smoothly, and is much less sensitive to small changes in the number of replicates. Surprisingly, this is not the case for GFOLD and edgeR sensitivity, which drops considerably when transitioning from 1 to 2 replicates, most likely due to the different statistical models employed with and without replicates. Although we varied the number of replicates without controlling the total number of reads as Liu et al. [15], our results strongly suggest that cost effectiveness metrics such as those proposed in [15] are likely to depend on to the specific method used for performing DE analysis. Therefore, the analysis method should be taken into account when using such a metric to guide the design of RNA-Seq experiments.

Effect of gene abundance

We also studied the effect of gene abundance on the IsoDE, GFOLD, and edgeR prediction accuracy. We selected the subset of genes that are expressed in at least one of the two RNA samples. We sorted these genes by the average of the gene's expression. We used the FPKM values predicted by IsoEM, the FPKM values predicted by GFOLD, and the number of uniquely mapped reads, for IsoDE, GFOLD, and edgeR, respectively. The genes were then divided into quintiles, for each method independently, where quintile 1 had the genes with the lowest expression levels, and quintile 5 had the genes with the highest expression levels. Sensitivity, PPV, and F-score where calculated for each quintile separately.

Figure 4 shows that, for results with both 1 and 6 replicates, sensitivity, PPV, and F-score of IsoDE are only slightly lower on genes with low expression levels compared to highly expressed genes (similar results are achieved for intermediate numbers of replicates and higher fold change thresholds). In contrast, GFOLD shows a marked difference in all accuracy measures for genes in the lower quintiles compared to those in the higher quintiles. The sensitivity of edgeR is also lower for genes expressed at low levels, however it's PPV is relatively constant across expression levels.

Conclusions

A practical bootstrapping based method, IsoDE, was developed for analysis of differentially expressed genes in RNA-Seq datasets. Unlike other existing methods, IsoDE is non-parametric, i.e., does not assume an underlying statistical distribution of the data. Experimental results on publicly available datasets both with and without replicates show that IsoDE has robust performance over a wide range of technologies, sequencing depths, and minimum fold changes. IsoDE performs particularly well on low coverage RNA-Seq datasets, at low fold change thresholds, and when no or very few replicates are available.

Availability

IsoDE has been implemented in Java and can be run on any platform with a Java virtual machine. The source code and installation instructions are available at http://dna.engr.uconn.edu/software/IsoDE/. A web-based calculator for computing the bootstrap support based on the desired number of bootstrap samples and significance level is available at http://dna.engr.uconn.edu/~software/cgi-bin/calc/calc.cgi.

Competing interests

The authors declare that they have no competing interests.

Author's contributions
SS implemented and tested Fisher's exact test, prepared the testing data used to generate the experimental results, ran Cuffdiff, GFOLD, and edgeR, and performed comparisons between these methods and IsoDE. SS also contributed to designing the algorithms and wrote part of the manuscript. YTT developed, implemented and tested the algorithms for IsoDE and wrote part of the manuscript. AZ contributed to designing the algorithms and the experiments, writing the manuscript and supervised the project. IM contributed to designing the algorithms and the experiments, writing the manuscript and supervised the project. All authors read and approved the final manuscript. . .

Acknowledgements
This work has been partially supported by the Agriculture and Food Research Initiative Competitive Grant No. 2011-67016-30331 from the USDA National Institute of Food and Agriculture and awards IIS-0916401 and IIS-0916948 from NSF, a Collaborative Research Grant from Life Technologies, and the Molecular Basis of Disease Area of Focus Georgia State University. . .

Author details
[1]Computer Science & Engineering Department, University of Connecticut, 06269 Storrs, CT, USA. [2]Computer Science Department, Georgia State University, 34 Peachtree str., 30303 Atlanta, GA, USA.

References
1. 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)
2. 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)
3. Wang, Z., Gerstein, M., Snyder, M.: Rna-seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics **10**(1), 57–63 (2009)
4. 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)
5. 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)
6. Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome Biol **11**(10), 106 (2010)
7. Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., Pachter, L.: Differential analysis of gene regulation at transcript resolution with rna-seq. Nature biotechnology **31**(1), 46–53 (2012)
8. Bi, Y., Davuluri, R.V.: Npebseq: nonparametric empirical bayesian-based procedure for differential expression analysis of rna-seq data. BMC bioinformatics **14**(1), 262 (2013)
9. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., *et al.*: Ncbi geo: archive for functional genomics data sets—10 years on. Nucleic acids research **39**(suppl 1), 1005–1010 (2011)
10. 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)
11. Efron, B., Tibshirani, R.: An Introduction to the Bootstrap. Macmillan Publishers Limited, ??? (1993)
12. 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)
13. Consortium, M.: The Microarray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nature Biotechnology **24**(9), 1151–1161 (2006)
14. Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D., Betel, D.: Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. Genome biology **14**(9), 95 (2013)
15. Liu, Y., Zhou, J., White, K.P.: Rna-seq differential expression studies: more sequence or more replication? Bioinformatics **30**(3), 301–304 (2014)
16. Nicolae, M., Mangul, S., Mandoiu, I.I., Zelikovsky, A.: Estimation of alternative splicing isoform frequencies from rna-seq data. Algorithms for Molecular Biology **6:9** (2011)
17. Bo Li, C.N.D.: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome (2011)
18. Reid, L.H.: Proposed methods for testing and selecting the ercc external rna controls. BMC genomics **6**(1), 1–18 (2005)
19. Yoav Benjamini, Y.H.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society, Series B **57**(1), 289–300 (1995)
20. 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), 25 (2009)
21. Lee, W.-P., Stromberg, M.P., Ward, A., Stewart, C., Garrison, E.P., Marth, G.T.: MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. PLoS ONE **9**(3), 90581 (2014)

Figures

Tables
Additional Files

	Ground truth		
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 gene expression

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%
	edgeR	73.03%	73.26%	95.56%	82.94%
	IsoDE-Match	**82.63%**	**87.46%**	83.70%	85.54%
	IsoDE-All	82.22%	87.17%	82.82%	84.94%
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%
	edgeR	**80.01%**	79.92%	92.07%	**85.57%**
	IsoDE-Match	78.98%	86.23%	84.62%	85.42%
	IsoDE-All	79.01%	**86.42%**	84.49%	85.44%
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%
	edgeR	**83.67%**	81.63%	91.17%	**86.13%**
	IsoDE-Match	82.04%	85.58%	85.19%	85.38%
	IsoDE-All	81.20%	**86.74%**	83.07%	84.87%

Table 2: Accuracy, sensitivity, PPV and F-Score in % for MAQC Illumina dataset and fold change threshold f of 1, 1.5, and 2. The number of bootstrap samples is $M = 200$ for IsoDE-Match and $M = 20$ for IsoDE-All, and bootstrap support was determined using the binomial model with significance level $\alpha = 0.05$.

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%
	IsoDE-Match	**81.75%**	**86.38%**	82.18%	**84.05%**
	IsoDE-All	81.46%	86.07%	82.13%	**84.05%**
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%
	IsoDE-Match	77.66%	83.94%	84.75%	**84.34%**
	IsoDE-All	**77.81%**	**84.13%**	84.45%	84.29%
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%
	IsoDE-Match	**81.75%**	85.81%	84.63%	**85.22%**
	IsoDE-All	81.61%	**86.28%**	84.13%	85.19%

Table 3: Accuracy, sensitivity, PPV and F-Score in % for Ion Torrent dataset and fold change threshold f of 1, 1.5, and 2. The number of bootstrap samples is $M = 200$ for IsoDE-Match and $M = 20$ for IsoDE-All, and bootstrap support was determined using the binomial model with significance level $\alpha = 0.05$.

Fold Change	Method	Accuracy %	Sensitivity %	PPV %	F-Score %
1	FisherTotal	34.01%	30.50%	95.63%	46.24%
	FisherHousekeeping	24.52%	20.12%	**94.74%**	33.38%
	GFOLD	55.62%	54.18%	92.11%	68.23%
	IsoDE-Match	75.33%	79.57%	77.41%	78.47%
	IsoDE-All	**78.85%**	**84.67%**	81.04%	**82.82%**
1.5	FisherTotal	48.18%	35.37%	89.81%	50.75%
	FisherHousekeeping	42.48%	24.86%	**97.74%**	39.63%
	GFOLD	62.19%	58.13%	85.39%	69.17%
	IsoDE-Match	64.09%	74.19%	72.52%	73.35%
	IsoDE-All	**72.85%**	**79.54%**	80.62%	**80.08%**
2	FisherTotal	57.96%	39.53%	85.43%	54.05%
	FisherHousekeeping	55.33%	29.30%	**97.67%**	45.08%
	GFOLD	69.05%	61.16%	83.49%	70.60%
	IsoDE-Match	67.15%	76.51%	70.30%	73.27%
	IsoDE-All	**75.18%**	**80.93%**	78.03%	**79.45%**

Table 4: Accuracy, sensitivity, PPV and F-Score in % for the First 454 dataset and fold change threshold f of 1, 1.5, and 2. The number of bootstrap samples is $M = 200$ for IsoDE-Match and $M = 20$ for IsoDE-All, and bootstrap support was determined using the binomial model with significance level $\alpha = 0.05$.

Data set	Sample	Reads Length	Mapped Reads*	Uniquely Mapped Reads*	Notes
SRX365211	MCF-7 Control	49	24.22	18.57	MCF-7 Control Replicate 1
SRX365212	MCF-7 Control	49	41.85	31.98	MCF-7 Control Replicate 2
SRX365213	MCF-7 Control	49	32.49	24.63	MCF-7 Control Replicate 3
SRX365214	MCF-7 Control	49	32.86	25.13	MCF-7 Control Replicate 4
SRX365215	MCF-7 Control	49	33.05	25.18	MCF-7 Control Replicate 5
SRX365216	MCF-7 Control	49	40.62	30.90	MCF-7 Control Replicate 6
SRX365217	MCF-7 Control	49	64.34	49.05	MCF-7 Control Replicate 7
SRX365204	MCF-7 E1	49	28.58	21.58	MCF-7 E2 Replicate 1
SRX365205	MCF-7 E2	49	23.71	18.26	MCF-7 E2 Replicate 2
SRX365206	MCF-7 E3	49	26.75	20.92	MCF-7 E2 Replicate 3
SRX365207	MCF-7 E4	49	24.95	18.64	MCF-7 E2 Replicate 4
SRX365208	MCF-7 E5	49	28.35	21.52	MCF-7 E2 Replicate 5
SRX365209	MCF-7 E6	49	27.08	20.80	MCF-7 E2 Replicate 6
SRX365210	MCF-7 E7	49	30.66	23.38	MCF-7 E2 Replicate 7
SRX003926	MAQC HBRR	36	6.80	4.8	MAQC Illumina dataset
SRX003927	MAQC UHRR	36	5.80	3.9	MAQC Illumina dataset
SRX002934	MAQC UHRR	253.4	0.52		MAQC First 454 dataset
SRX002935	MAQC HBRR	251.1	0.50		MAQC First 454 dataset
SRX002932	MAQC UHRR	253.6	0.54		MAQC Second 454 dataset
SRX002933	MAQC HBRR	252.0	0.78		MAQC Second 454 dataset
DID-143-282	MAQC HBRR	94.2	1.22		MAQC ION Torrent
LUC-140-265	MAQC HBRR	97.7	1.14		MAQC ION Torrent
GOG-139-281	MAQC HBRR	93.3	1.21		MAQC ION Torrent
LUC-141-267	MAQC HBRR	95.1	1.04		MAQC ION Torrent
POZ-124-266	MAQC HBRR	95.5	1.07		MAQC ION Torrent
DID-143-283	MAQC UHRR	99.8	1.37		MAQC ION Torrent
GOG-140-284	MAQC UHRR	101.1	1.45		MAQC ION Torrent
POZ-125-268	MAQC UHRR	96.8	1.10		MAQC ION Torrent
POZ-126-269	MAQC UHRR	102.1	1.29		MAQC ION Torrent
POZ-127-270	MAQC UHRR	99.8	1.62		MAQC ION Torrent

Table 5: Datasets used in the experimental analysis and their mapping statistics. Number of Uniquely mapped reads is specified for the datasets analyzed by edgeR. *In million reads

Fold Change	Method	Accuracy %	Sensitivity %	PPV %	F-Score %
1	FisherTotal	34.01%	30.49%	95.63%	46.24%
	FisherHousekeeping	24.53%	20.12%	**97.74%**	33.38%
	GFOLD	55.62%	54.18%	92.11%	68.23%
	IsoDE-Match	**78.39%**	**83.13%**	79.56%	**81.30%**
	IsoDE-All	78.10%	82.66%	79.23%	80.91%
1.5	FisherTotal	48.18%	35.37%	89.81%	50.75%
	FisherHousekeeping	42.48%	24.86%	**97.74%**	39.63%
	GFOLD	62.19%	58.13%	85.39%	69.17%
	IsoDE-Match	**71.09%**	**79.35%**	79.20%	**79.27%**
	IsoDE-All	70.36%	78.59%	79.04%	78.81%
2	FisherTotal	57.96%	39.53%	85.43%	54.05%
	FisherHousekeeping	55.33%	29.30%	**97.67%**	45.08%
	GFOLD	69.05%	61.16%	83.49%	70.60%
	IsoDE-Match	**74.60%**	**80.47%**	78.10%	**79.27%**
	IsoDE-All	71.97%	79.30%	75.61%	**77.41%**

Table 6: Accuracy, sensitivity, PPV and F-Score in % for the Second 454 dataset and fold change threshold f of 1, 1.5, and 2. The number of bootstrap samples is $M = 200$ for IsoDE-Match and $M = 20$ for IsoDE-All, and bootstrap support was determined using the binomial model with significance level $\alpha = 0.05$.

Fold Change	Method	Accuracy %	Sensitivity %	PPV %	F-Score %
1	FishersTotal	49.05%	46.44%	97.09%	62.83%
	FishersHousekeeping	40.88%	37.62%	**98.38%**	54.42%
	FisherERCC	52.55%	51.70%	88.83%	65.36%
	GFOLD	59.27%	59.29%	91.41%	71.92%
	IsoDE-All	**79.12%**	**83.75%**	79.91%	**81.78%**
1.5	FisherTotal	60.29%	53.35%	90.29%	67.07%
	FisherHousekeeping	57.08%	45.31%	**96.34%**	61.64%
	FisherERCC	57.96%	56.02%	82.07%	66.59%
	GFOLD	67.01%	65.58%	87.72%	75.05%
	IsoDE-All	**72.55%**	**80.50%**	80.34%	**80.42%**
2	FisherTotal	69.05%	59.53%	86.78%	70.62%
	FisherHousekeeping	68.90%	53.49%	**94.26%**	68.25%
	FisherERCC	66.42%	60.23%	80.93%	69.07%
	GFOLD	72.85%	66.51%	86.93%	75.36%
	IsoDE-All	**76.64%**	**81.40%**	81.02%	**81.21%**

Table 7: Accuracy, sensitivity, PPV and F-Score in % for ION Torrent pair HBRR: LUC-140_265 and UHRR: POZ-126_269, with fold change threshold f of 1, 1.5, and 2. The number of bootstrap samples for IsoDE-All is $M = 20$, and bootstrap support was determined using the binomial model with significance level $\alpha = 0.05$.

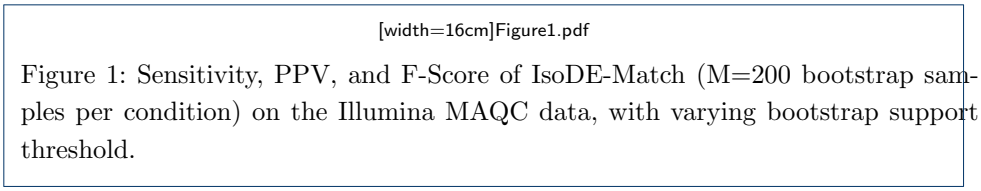


Figure 1: Sensitivity, PPV, and F-Score of IsoDE-Match (M=200 bootstrap samples per condition) on the Illumina MAQC data, with varying bootstrap support threshold.

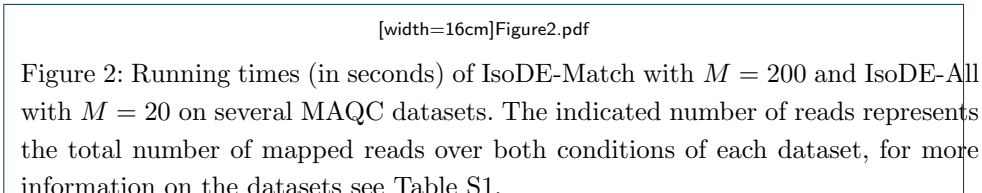


Figure 2: Running times (in seconds) of IsoDE-Match with $M = 200$ and IsoDE-All with $M = 20$ on several MAQC datasets. The indicated number of reads represents the total number of mapped reads over both conditions of each dataset, for more information on the datasets see Table S1.

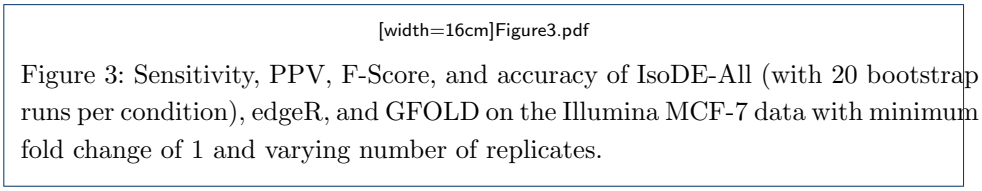


Figure 3: Sensitivity, PPV, F-Score, and accuracy of IsoDE-All (with 20 bootstrap runs per condition), edgeR, and GFOLD on the Illumina MCF-7 data with minimum fold change of 1 and varying number of replicates.

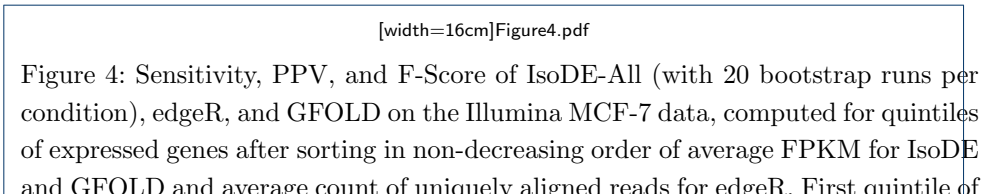


Figure 4: Sensitivity, PPV, and F-Score of IsoDE-All (with 20 bootstrap runs per condition), edgeR, and GFOLD on the Illumina MCF-7 data, computed for quintiles of expressed genes after sorting in non-decreasing order of average FPKM for IsoDE and GFOLD and average count of uniquely aligned reads for edgeR. First quintile of edgeR had 0 differentially expressed genes according to the ground truth (obtained by using all 7 replicates).

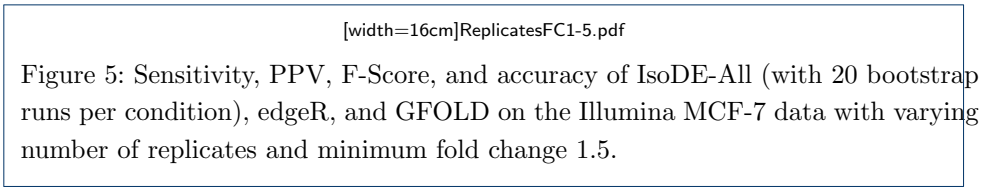


Figure 5: Sensitivity, PPV, F-Score, and accuracy of IsoDE-All (with 20 bootstrap runs per condition), edgeR, and GFOLD on the Illumina MCF-7 data with varying number of replicates and minimum fold change 1.5.

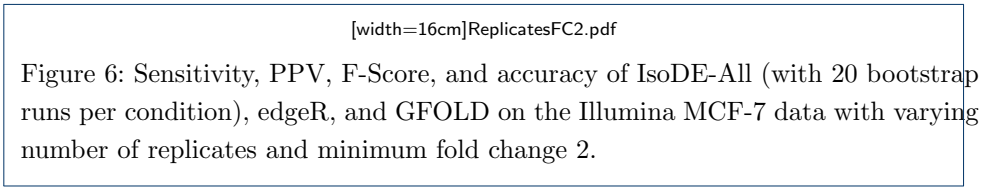


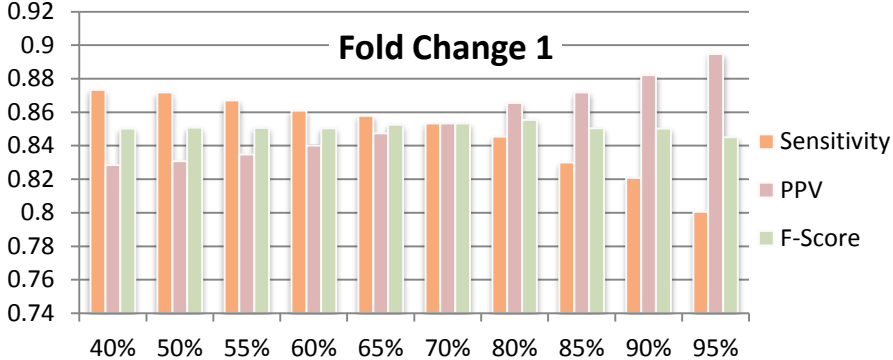
Figure 6: Sensitivity, PPV, F-Score, and accuracy of IsoDE-All (with 20 bootstrap runs per condition), edgeR, and GFOLD on the Illumina MCF-7 data with varying number of replicates and minimum fold change 2.

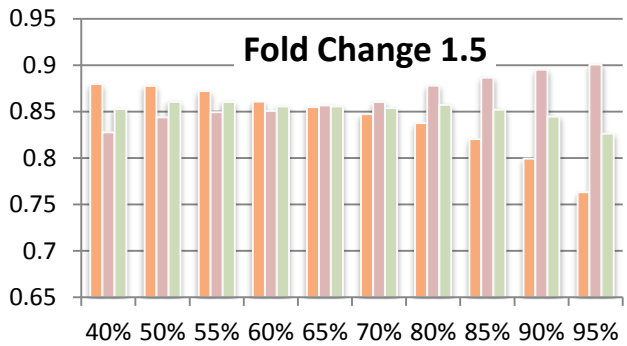
Fold Change	Method	Accuracy %	Sensitivity %	PPV %	F-Score %
1	FishersTotal	51.09%	48.76%	96.04%	64.68%
	FishersHousekeeping	46.42%	43.65%	**97.58%**	60.32%
	FisherERCC	55.04%	54.33%	88.86%	67.43%
	GFOLD	60.44%	60.84%	83.09%	70.24%
	IsoDE-All	**79.56%**	**84.37%**	80.50%	**82.39%**
1.5	FisherTotal	62.34%	56.02%	90.43%	69.19%
	FisherHousekeeping	61.61%	52.20%	**94.46%**	67.24%
	FisherERCC	60.44%	57.17%	85.92%	68.66%
	GFOLD	64.09%	62.91%	82.87%	71.52%
	IsoDE-All	**74.60%**	**80.69%**	82.10%	**81.39%**
2	FisherTotal	69.05%	60.47%	85.81%	70.94%
	FisherHousekeeping	70.36%	58.84%	**90.03%**	71.17%
	FisherERCC	67.30%	60.00%	82.96%	69.64%
	GFOLD	72.55%	65.12%	86.96%	74.45%
	IsoDE-All	**76.50%**	**79.77%**	81.28%	**80.52%**

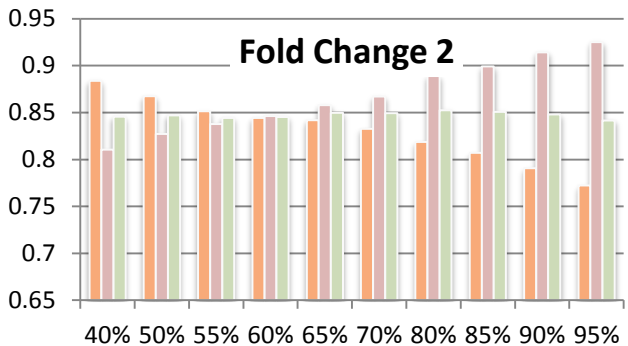
Table 8: Accuracy, sensitivity, PPV and F-Score in % for ION Torrent pair HBRR: GOG-139_281 and UHRR: POZ-127_270, with fold change threshold f of 1, 1.5, and 2. The number of bootstrap samples for IsoDE-All is $M = 20$, and bootstrap support was determined using the binomial model with significance level $\alpha = 0.05$.

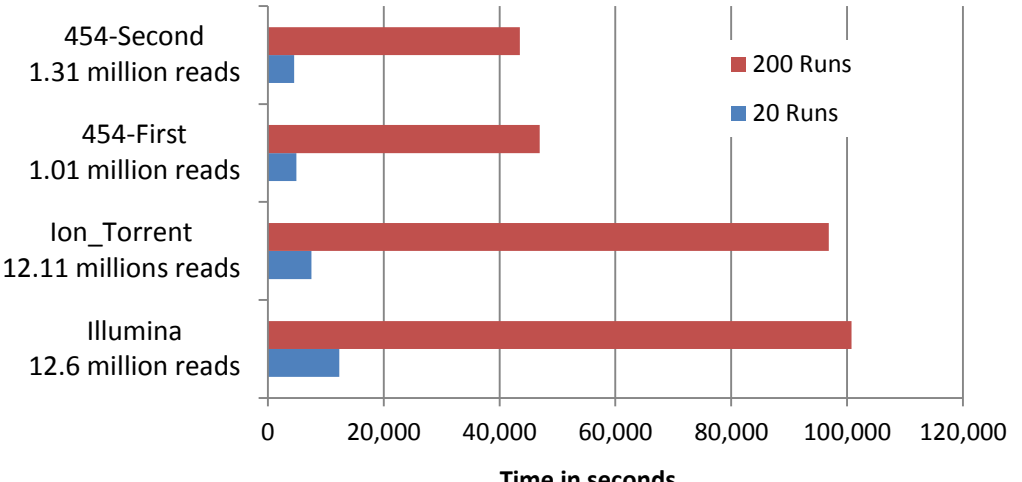
Number of Replicates	Rep1	Rep2	Rep3	Rep4	Rep5	Rep6	Rep7	Bootstrap Samples Per Condition
1	20							20
2	10	10						20
3	7	7	6					20
4	5	5	5	5				20
5	4	4	4	4	4			20
6	4	4	3	3	3	3		20
7	3	3	3	3	3	3	2	20

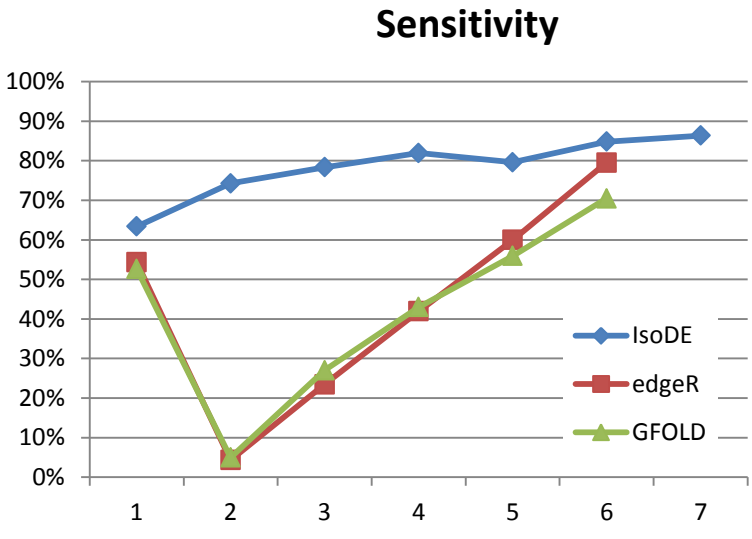
Table 9: IsoDE setup for experiments with replicates. IsoDE experiments on the MCF-7 dataset was performed as follows. First we generated, for each of the 7 replicates of each condition 20, 10, 6, 5, 4, 3, respectively 2 bootstrap samples. We then used subsets of these bootstrap samples as input for IsoDE to perform DE analysis with varying number of replicates and a fixed total number $M = 20$ of bootstrap samples per condition. In experiment 1 we used 20 bootstrap samples from first replicate of each condition, in experiment 2 we used 10 bootstrap samples for each of the first 2 replicates of each condition, and so on.

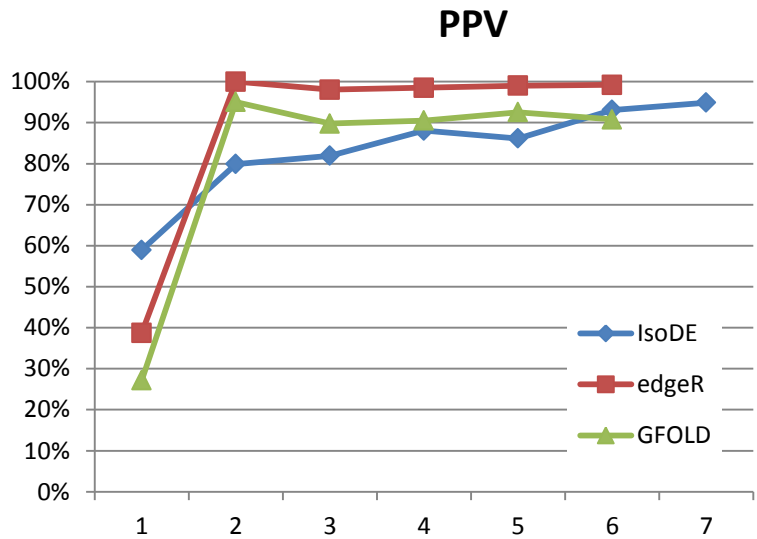

Fold Change 1
0.92
0.9
0.88
0.86
0.84
0.82
0.8
0.78
0.76
0.74
40% 50% 55% 60% 65% 70% 80% 85% 90% 95%
Sensitivity
PPV
F-Score

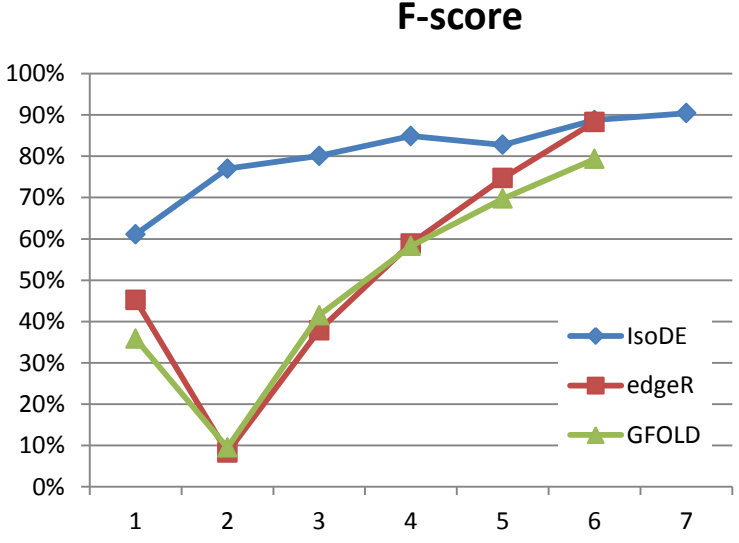

Fold Change 1.5
0.95
0.9
0.85
0.8
0.75
0.7
0.65
40% 50% 55% 60% 65% 70% 80% 85% 90% 95%

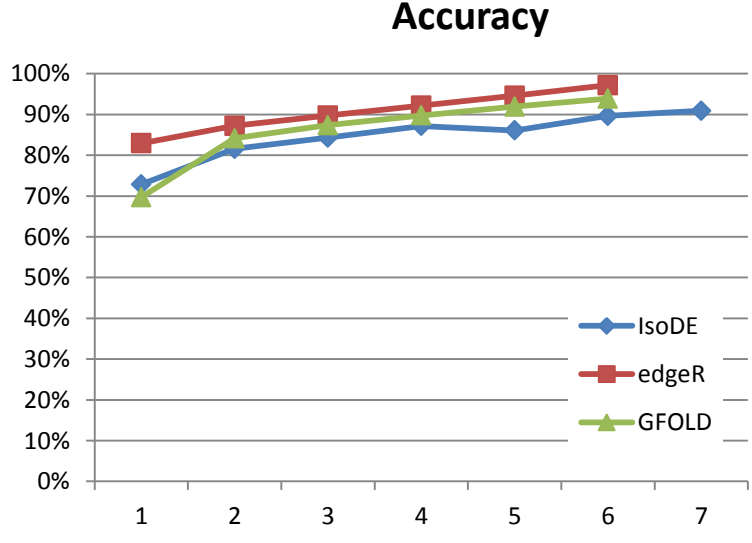

Fold Change 2
0.95
0.9
0.85
0.8
0.75
0.7
0.65
40% 50% 55% 60% 65% 70% 80% 85% 90% 95%


454-Second
1.31 million reads
454-First
1.01 million reads
Ion_Torrent
12.11 millions reads
Illumina
12.6 million reads
200 Runs
20 Runs
0
20,000
40,000
60,000
80,000
100,000
120,000
Time in seconds

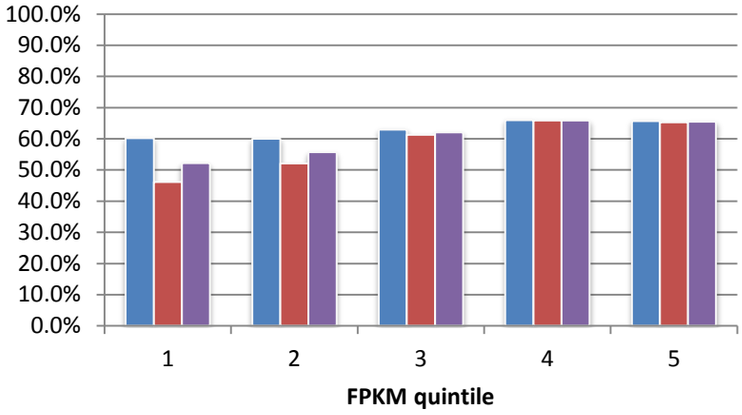

Sensitivity
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
IsoDE
edgeR
GFOLD


PPV
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
IsoDE
edgeR
GFOLD

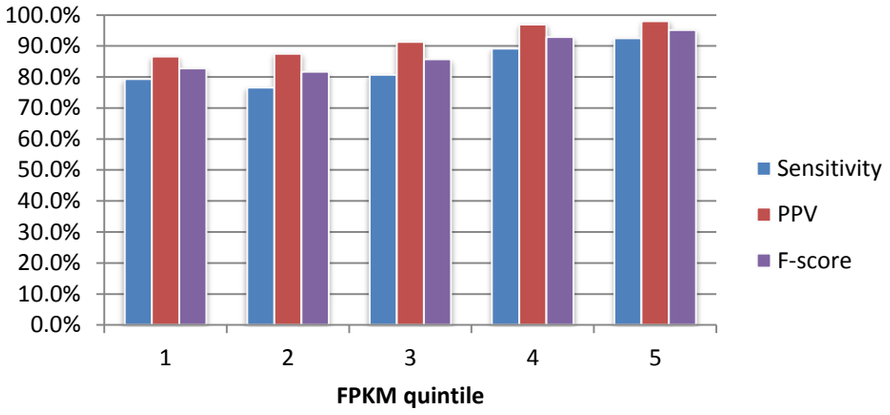

F-score
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
IsoDE
edgeR
GFOLD


Accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
IsoDE
edgeR
GFOLD

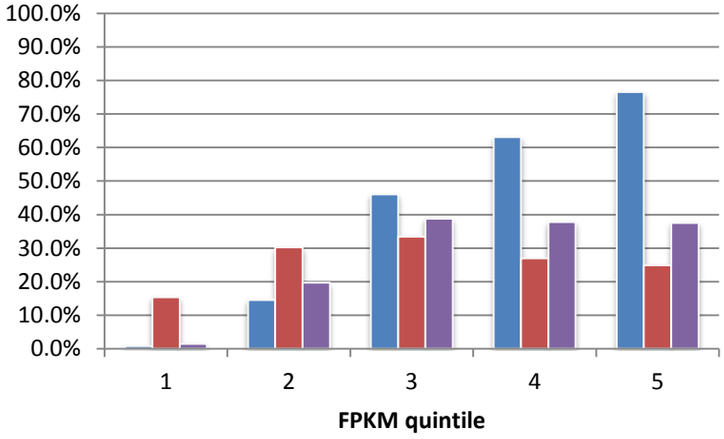
IsoDE - 1 replicate - FC = 1



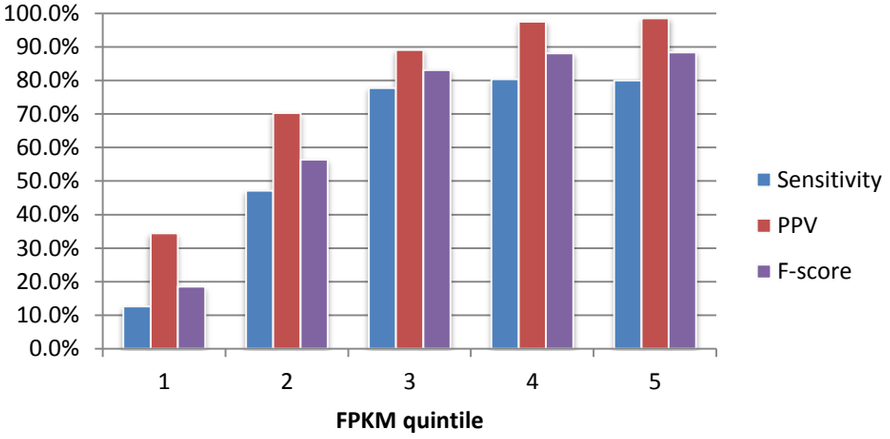
IsoDE - 6 replicates - FC = 1



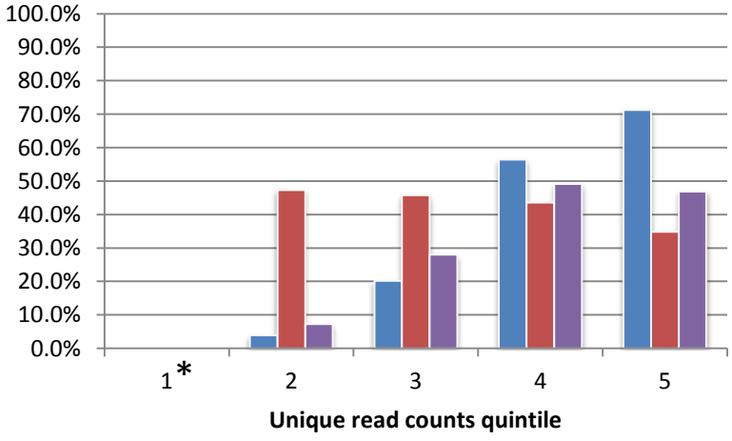
GFOLD - 1 replicate - FC = 1



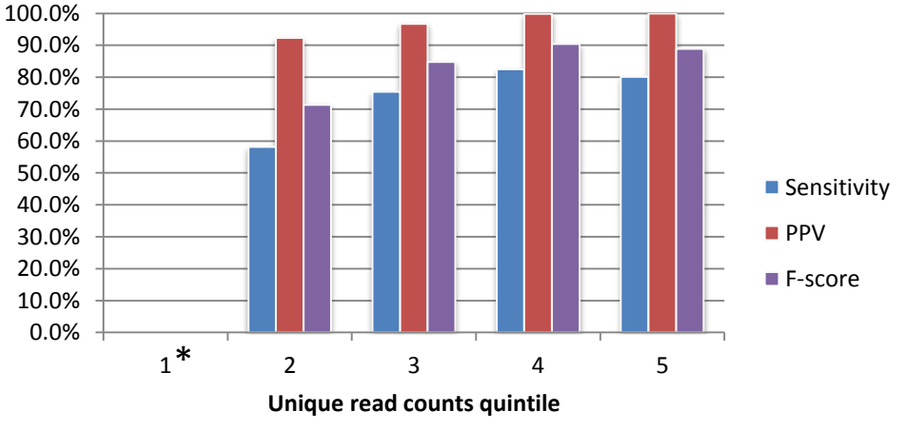
GFOLD - 6 replicates - FC = 1

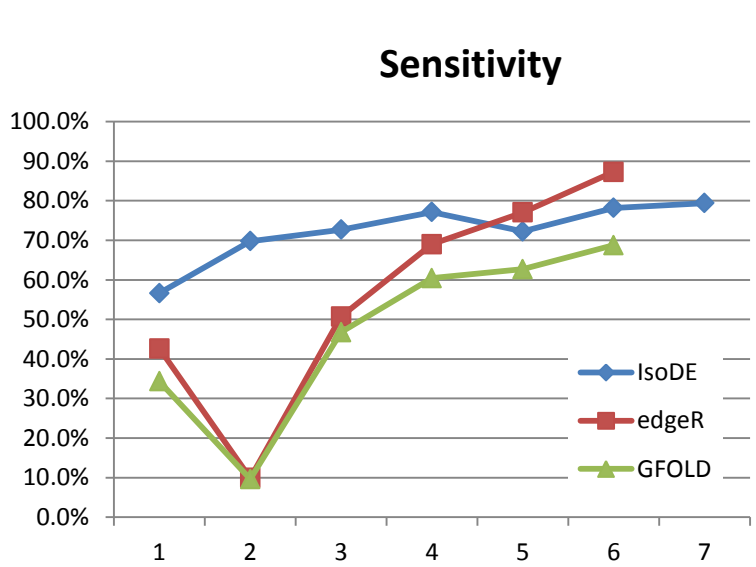


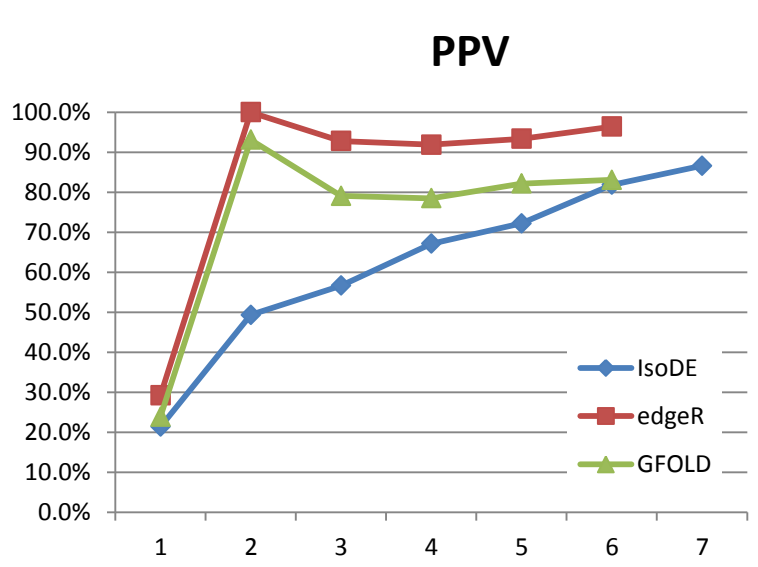
edgeR - 1 replicate - FC = 1

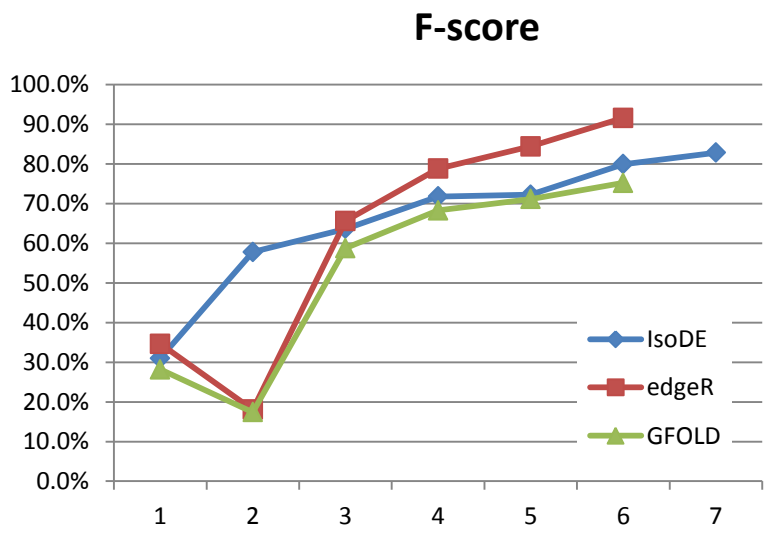


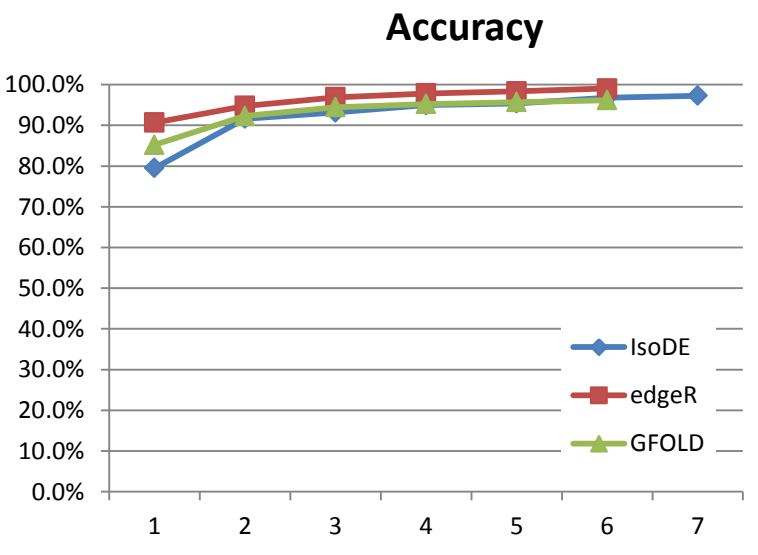
edgeR - 6 replicates - FC = 1



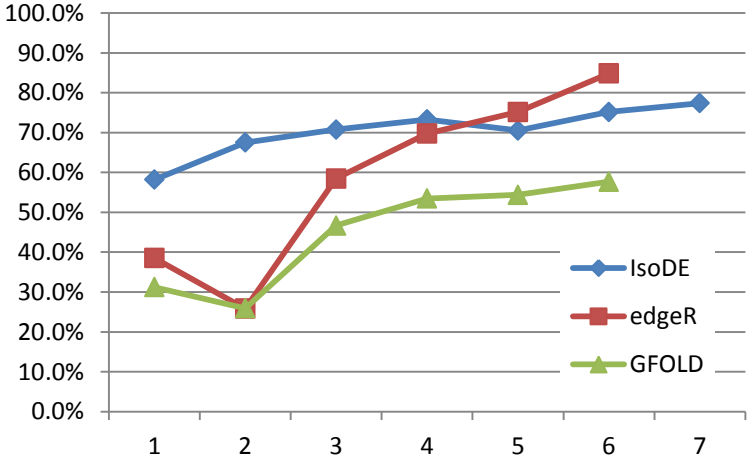

Sensitivity
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
1
2
3
4
5
6
7
IsoDE
edgeR
GFOLD


PPV
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
1
2
3
4
5
6
7
IsoDE
edgeR
GFOLD

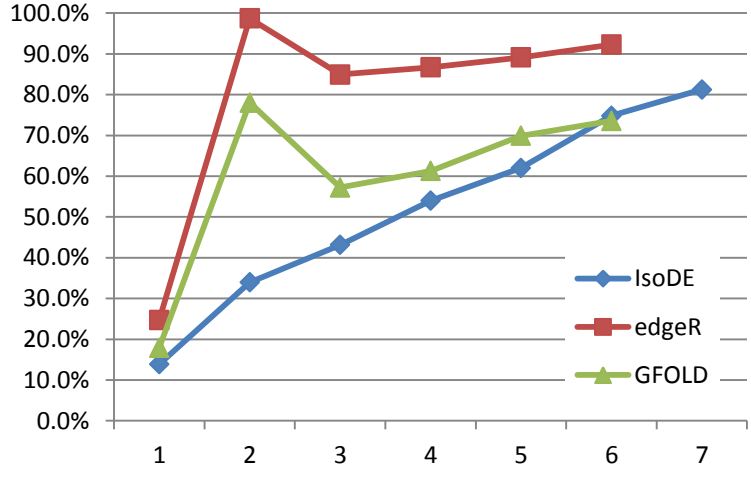

F-score
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
1
2
3
4
5
6
7
IsoDE
edgeR
GFOLD


Accuracy
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
1
2
3
4
5
6
7
IsoDE
edgeR
GFOLD

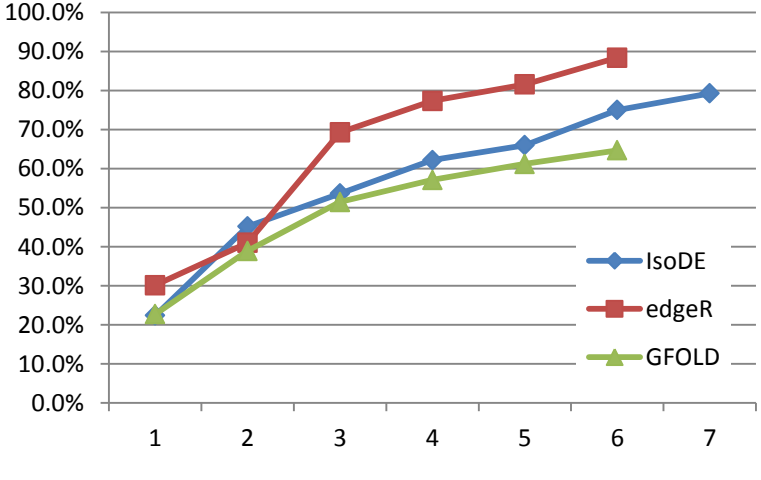
Sensitivity



PPV



F-score



Accuracy