tr
Var
median

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–11

iDEG: a single-subject method for assessing gene differential expression from two transcriptomes of an individual

Qike Li [1] , Colleen Kenost [2] , Yves [3]

[1]Affiliation not available
[2]Colleen Kenost [3]Yves

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Single-subject RNA-Sequencing (RNA-Seq) analysis is a powerful precision medicine tool for unveiling individual disease mechanisms. Due to the scarcity of relevant tissue samples and the cost of high-throughput technologies, the availability of replicates for a single subject is often impractical. This constraint prohibits the use of most conventional statistical techniques since replicates are typically needed to estimate data variability and make inferences. We propose the iDEG method to identify individualized Differentially Expressed Genes from two conditions of a subject, each sampled once: a baseline sample (e.g., unaffected tissue) vs. a case sample (e.g., cancer). iDEG gathers information across different genes from the same individual while strategically bypassing the requirement of replicates per condition to make valid inferences. The main idea of iDEG is to transform RNA-Seq data such that, under the null hypothesis, differences of transformed expression counts follow the same distribution with a constant variance. This transformation enables modeling all genes with a two-group mixture model, from which the probability of differential expression for each gene is then estimated by an empirical Bayes approach with a local false discovery rate control. Our extensive numerical studies demonstrate iDEG's superior performance compared to existing methods under a variety of scenarios. Finally, iDEG is applied to a triple negative breast cancer single-subject dataset in which a personal set of differentially expressed genes are identified.

Availability: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text

Contact: yves@email.arizona.edu

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

\firstpage
1
\address
[1]Statistics GIDP, University of Arizona, Tucson, AZ, U.S.A
[2]Center for Biomedical Informatics and Biostatistics, University of Arizona, Tucson, AZ, U.S.A.
[3]Department of Medicine, University of Arizona, Tucson, AZ, U.S.A
[4]Department of Math, University of Arizona, Tucson, AZ, U.S.A
[5]University of Arizona Cancer Center, University of Arizona, Tucson, AZ, U.S.A.
[6]BIO5 Institute, University of Arizona, Tucson, AZ, U.S.A.
*
\corresp
*To whom correspondence should be addressed.
\history
Received on XXXXX; revised on XXXXX; accepted on XXXXX
\editor
Associate Editor: XXXXXXX

INTRODUCTION

Single-subject Analysis for RNA-Seq Data

Precision medicine, also known as personalized medicine or individualized medicine, aims to deliver "the right treatments, at the right time, every time to the right person" Kaiser (2015). The conventional one-size-fits-all drug development approach to the heterogeneous human population has led to the ten top-grossing USA drugs being ineffective for more than 75% of users Schork (2015). In contrast, precision medicine tailors the optimal treatment to individuals, and clinical trial designs are moving from cohort-based to single-subject trials Schork (2015). The success of precision medicine hinges on identifying personal disease mechanisms Topol (2014) to optimize disease treatment regimens based on an individual's biology (e.g., response to stimuli, genomic profile, and baseline risk among other factors).

Single-subject RNA sequencing (RNA-Seq) analysis considers one patient at a time, with the goal of revealing the altered transcriptomic mechanisms, for example, those associated to a disease state of this patient. Compared to traditional cohort-based analysis, the major challenge of single-subject RNA-Seq analysis is estimating the variance of gene expression levels when there are no replicates for each subject - i.e., the single-subject, single-transcriptome per condition. Variance estimation is a central question in RNA-Seq analysis, as it plays a key role in identifying altered mechanisms such as differentially expressed genes (DEGs). Conventional statistics estimate the variance from biological or technical replicates. However, obtaining transcriptome replicates is difficult due to (i) limited tissue availability, (ii) the risks associated with invasive

tissue-sampling procedures, and (iii) general costs and inefficiencies with the current technology. As access to replicates in single subjects is compromised for the above-mentioned reasons and in order to advance precision medicine, the field requires novel methods designed to handle single-subject transcriptome analyses.

A Motivating Study

Breast cancer is one of the most common cancers with ∼500,000 deaths worldwide each year Wild and Stewart (2014). Cohort-based analyses have yielded valuable insights into providing personalized treatments by classifying breast cancers into four major subtypes Sørlie *et al.* (8418). However, no two cancers are alike, as significant heterogeneity is present within each subtype Cancer Genome Atlas Network (2012). Furthermore, minorities are underrepresented in most clinical trials, and, therefore, knowledge derived from such clinical trials may not be applicable to diverse populations. For example, triple negative breast cancer (TNBC) is a subtype of breast cancer that has poor prognosis and considerable heterogeneity, as well as disproportionately affects women from African origin Dietze (2015). In The Cancer Genome Atlas (TCGA) project, which collected RNA-Seq data on 1092 breast cancer patients, matched tumor/healthy samples were available from only two African American (AA) patients who differed remarkably in age, stage of tumor, survival, and other key features. In this case, single-subject RNA-Seq analysis would be more appropriate for discovering individual-specific DEGs and for identifying the best therapeutic options. Our study is specifically motivated by this single-subject RNA-Seq dataset downloaded from TCGA (Table **??**).

TNBC example (RNA-Seq single-subject dataset). The expression of the first ten genes in alphabetical order among 20,501 gene expression measurements, which are mapped to gene symbols for both tumor and surrounding healthy tissue collected from an African American female (subject TCGA-GI-A2C9) exhibiting TNBC. The second and third column display the mRNA counts of her healthy sample and the ones of her tumor sample. The last three columns - Absolute Difference, Fold Change (FC), and indicator of FC $\geq$ 3 or $\frac{1}{FC} \geq 3$ - illustrate the general complexity of working with count data and the caution one must proceed with when developing methods for both lowly and highly expressed genes. Using a simple heuristic of FC ≥ 3 or $\frac{1}{FC} \geq 3$ to label a DEG, we see two potential extreme cases of misclassifying a gene by assuming that genes of different orders of magnitude present the same behavior. Gene A2M, for example, has an absolute difference of 17,560 and $\frac{1}{FC} = 2.5$, which could be a potential prime candidate for a down-regulated DEG. Even though A4GNT has a FC=5, it may not be a DEG since there tends to be more noise than signal at such low levels of expression. Note, single-subject RNA-Seq analysis compare isogenic tissues of the same subject, and isogenic refers to identical genomes as in tissues of the same subject, cell lines, or highly inbred animal models (e.g., mice strains), while heterogenic conditions are observed between individuals with distinct genomes (e.g., most human beings).

Gene	Healthy	Tumor	Absolute Difference	Fold Change(FC)	FC ≥ 3 or $\frac{1}{FC} \geq 3$
A1BG	72	92	20	1.28	0
A1CF	0	1	1	NaN	NA
A2BP1	2	0	2	0	NA

Gene	Healthy	Tumor	Absolute Difference	Fold Change(FC)	$FC \geq 2$ or $\frac{1}{FC} \geq 2$
A2LD1	71	127	56	1.79	0
A2ML1	12	773	761	64.42	1
A2M	29385	11825	17560	0.4	0
A4GALT	891	871	20	0.98	0
A4GNT	5	1	4	0.2	1
AAA1	0	0	0	NaN	NA
AAAS	460	414	46	0.9	0

Legend. NaN:not defined. NA: not applicable.

This study aims to discover which genes have significantly differential expressions between tumor and normal samples *for each single patient* . However, the main challenge lies in each gene being measured only once under each condition. In single-subject analyses, conventional analytics are either infeasible or underpowered to detect changes. Therefore, we propose a novel strategy, iDEG (Identifying individualized sets of Differentially Expressed Genes), to overcome this challenge for identifying important genes effectively. The new methodology is then applied to this TCGA dataset and the results are presented in Section .

DEG Identification for Single-subject Analysis

The random variables Y_{g1} and Y_{g2} are used to denote the expression counts of gene g under Condition 1 (e.g., normal) and Condition 2 (e.g., tumor). Furthermore, assume $\mu_{g1} = E(Y_{g1})$ and $\mu_{g2} = E(Y_{g2})$, their respective mean expression levels. In single-subject analyses, there is only one sample y_{g1} and only one sample y_{g2} observed for Y_{g1} and Y_{g2}, respectively. The goal is to identify genes whose mean expression is different between the two conditions, i.e., $\mu_{g1} \neq \mu_{g2}$, for each single subject.

Although there is a body of literature concerning methods for identifying DEGs, very few methods have been developed to identify DEGs without transcriptome replicates. Typically, when no replicates are available, investigators compare an heuristic cutoff value to the absolute difference $|y_{g2} - y_{g1}|$ or the fold change y_{g2}/y_{g1}, and genes exceeding the cutoff value are declared differentially expressed. The cutoff is usually chosen based on the emipirical experience. \Citeauthorwang-2009-degseq (\citeyearwang-2009-degseq) developed DEGseq, which assumes the expression counts follow a binomial distribution. Based on the binomial distribution, they used a normal distribution to approximate the distribution of the $\log_2$ fold change ($\log_2 Y_{g1} - \log_2 Y_{g2}$) at a given expression intensity ($\log_2 Y_{g1} + \log_2 Y_{g2}$) and calculated a Z-score for each gene. However, DEGSeq is not designed to model over-dispersed count data due to the binomial distribution assumption. Anders and Huber (2010) proposed DESeq to discover DEGs with small sample sizes. When neither condition has replicates, DESeq is still applicable but has low power and a high false negative rate. It assumes that most genes are non-differentially expressed and estimates a mean-variance relationship by treating two samples as if they are replicates. Another popular method, edgeR Robinson and Smyth (2007), assumes RNA-Seq data follow a negative binomial distribution whose variance is determined only by the value of dispersion with a given mean. Without replicates, edgeR assigns the same value to the dispersion parameter of all genes and conducts a negative binomial (NB) exact test to compute p-values. Moreover, the value of dispersion is predetermined based on the investigator's biological knowledge rather than estimated from the data. Therefore, edgeR is not reliable when the assumption of a constant dispersion across genes is invalid or the predetermined value of the dispersion is inaccurate. Overall, there appears to be a lack of work in the literature on individualized DEG identification for single-subject, single-sample RNA-Seq analyses, which can hamper advances in personalized medicine.

In this work, we propose a novel method, called iDEG, to identify individualized Differentially Expressed Genes without requiring transcriptome replicates for either condition. iDEG first applies an appropriate variance-stabilizing transformation (VST) technique to RNA-Seq data such that, under null hypotheses, every gene's difference between two transformed expression counts approximately follows the same normal distribution with mean zero and a constant variance. This bypasses the estimation of variance for each gene and resolves the constraint of no replicates. Furthermore, iDEG models gene differences using a two-group mixture model and then estimates the probability of differential expression for each gene via empirical Bayes approach. The two groups in the mixture model correspond to differentially and non-differentially expressed genes, and an empirical null distribution is computed from the data.

In practice, investigators sometimes encounter the problem of unequal library sizes—the total starting material (input RNA) sequenced for one transcriptome is more than that for the other transcriptome, i.e., $\mathrm{E}(Y_{gd}) = k_d \mu_{gd}$ for $d = 1, 2$, where k_d is the library size for samples under condition d and $k_1 \neq k_2$. Then, under null hypothesis $\mu_{g1} = \mu_{g2}$, $\mathrm{E}(Y_{g1}) \neq \mathrm{E}(Y_{g2})$ due to the unequal library sizes. This makes the observed expression counts under two conditions not directly comparable, requiring an extra data normalization step before identifying DEGs. We first develop iDEG for equal library sizes and then extend it to unequal library sizes.

The rest of this article is organized as follows. Section **??** proposes the iDEG procedure for RNA-Seq data under the framework of Poisson distribution. Section **??** generalizes the iDEG for overdispersion expression counts for the Negative Binomial distribution. A practical issue of unequal library sizes is addressed in Section . Section describes the computational algorithm and implementation of iDEG. Extensive numerical studies are shown in Section to illustrate the performance of iDEG and compare it with existing methods. Section demonstrates the robustness of iDEG when model assumptions are violated. Section applies iDEG to the TNBC dataset described in Section . A final discussion is given in Section **??**.

UNEQUAL LIBRARY SIZES

The problem of unequal library sizes is commonly encountered in practice. If the total starting material (input RNA) sequenced for one transcriptome is different from that of the other transcriptome, then the observed expression counts under two conditions are not directly comparable. A normalization step is necessary prior to testing $E(Y_{g1}) = E(Y_{g2})$ for any g.

Normalizing Poisson Data

For Poisson distribution, unequal library sizes are accounted for by normalizing the data to reads per million (RPM; Mortazavi and et al 2008). The normalized gene count Y^*_{gd} is given by $Y^*_{gd} =$

Figure 1: Panel A depicts the raw difference $D_g = Y_{g_1} - Y_{g_2}$ for $20,000$ genes, suggesting that the variance of D_g increases as the mean μ_g increases; hence, there is no uniform cutoff to differentiate DEGs and null genes. Panel B illustrates that, for null genes, VST makes the variance of $D_g^* = h_{Pois}(Y_{g1}) - h_{Pois}(Y_{g2})$ constant regardless of their expression mean μ_g. Panel C illustrates the marginal distribution and the empirical null distribution computed for calculating *fdr*; where the purple solid line represents marginal density of Z_g, scaled to overlay the histogram; the orange dashed line displays the empirical null distribution of Z_g, and the two triangles are located at the decision boundary for calling DEGs. Panel D represents the probability of a gene being null given z_g (the solid curve), and the red dashed line displays the cutoff (local fdr ≤ 0.2) for defining a DEG.

$\frac{Y_{gd}}{\sum_{g=1}^{G} Y_{gd}} \times 10^6$ for all g, and $d = 1, 2$. After normalization, one can test $E(Y_{g1}^*) = E(Y_{g2}^*)$. Web Figure 1 (Panels A and B) of the supplementary material illustrates the effect of normalization. Before normalization, the median of expression levels in one transcriptome is far away from that of the other transcriptome; after normalization, the two medians are approximately at the same level and comparable.

Since the normalized data no longer follow a Poisson distribution, the transformation $h_{Pois}(\cdot)$ in (**??**) cannot be directly applied to Y_{gd}^*. Therefore, a different transformation for Y_{gd}^* is needed to stabilize variance. Recall that if $Y \sim \text{Poisson}(\mu)$, then $\sqrt{Y} \dot{\sim} \text{N}(\sqrt{\mu}, \frac{1}{4})$ Anscombe (1948). Using this fact, we propose the transformation $\sqrt{Y_{gd}^*}$, which is shown to approximately follow a normal distribution with a constant variance across all the genes in Corollary 1.

*A*ssume $Y_{gd} \sim \text{Poisson}(\mu_{gd})$ for $g = 1, \ldots, G; d = 1, 2$, and they are all independent. Denote the library size for samples under condition d by k_d. Then

$$\sqrt{Y_{gd}^*} \dot{\sim} \text{N}(\tilde{\mu}_{gd}, \tilde{\sigma}_d^2), \quad g = 1, \ldots, G; d = 1, 2,$$

where

$$\tilde{\mu}_{gd} = \sqrt{\frac{\mu_{gd}}{\sum_{g=1}^{G} \mu_{gd}}} \times 10^3, \;\; \tilde{\sigma}_d^2 = \frac{1}{4} \frac{1}{k_d \sum_{g=1}^{G} \mu_{gd}} \times 10^6.$$

Corollary **??** indicates $\tilde{\sigma}_d^2$ does not depend on g. The proof is given in the Web Appendix A. Therefore, under the null hypothesis, we have

$$D_g^* = \sqrt{Y_{g1}^*} - \sqrt{Y_{g2}^*} \dot{\sim} N(0, \sqrt{\tilde{\sigma}_1^2 + \tilde{\sigma}_2^2}), \quad \forall g \in \overline{\mathcal{G}}.$$

It is noted that D_g^* follows a normal distribution with a common variance $\tilde{\sigma}_1^2 + \tilde{\sigma}_2^2$ across all of the null genes. Following this, the proposed iDEG procedure can be applied next.

Normalizing Data from Negative Binomial

When RNA-Seq data follow the NB distribution, we propose applying a quantile adjustment procedure by Robinson and Smyth (\citeyearrobinson-2007-small-sampl) for normalization. Specifically, this technique adjusts the observed expression counts up if the library size is below the geometric mean and vice versa. For the null genes, this procedure creates pseudo-data that follow approximately an identical NB distribution. In Figure S1, the bottom row

shows roughly equal library sizes of the pseudo-data after normalization. Since the pseudo-data follow NB, the VST $h_{nb}(\cdot)$ in (**??**) can be applied, which is then followed by the iDEG for NB data (Section **??**).

COMPUTATIONAL ALGORITHMS

Sections and describe the iDEG algorithms for Poisson and Negative Binomial distributed RNA-Seq data, respectively. In practice, investigators need to specify the distribution assumption according to their understanding of the data. The Negative Binomial distribution is more general, and its limiting case is the Poisson distribution when the dispersion parameter δ_g goes to zero. We have developed an R package iDEG, available at https://github.com/QikeLi/iDEG.

iDEG Algorithm for Poisson RNA-Seq data described in Table ??

iDEG algorithm

For negative binomial RNA-Seq data:

- **Step 0** Normalize the data (*for unequal library sizes only*).
- **Step 1.** Group genes into windows based on their gene expression levels as in (**??**).
- **Step 2.** Compute $\hat{\bar{\mu}}_w$ and $\hat{\bar{\sigma}}^2_w$ for each window w, and obtain a "raw" estimate of δ_g.
- **Step 3.** Obtain a "refined" estimate of δ_g by fitting a smoothing spline.
- (**Step 3'.** Alternatively, if a constant dispersion is more appropriate, fit a linear regression model (**??**) to estimate the dispersion value $\hat{\delta}_0$.)
- **Step 4.** Apply the VST h_{nb} (**??**) to each gene expression count.
- **Step 5.** Compute the standardized summary statistics Z_g for each gene.
- **Step 6.** Estimate the local false discovery rate locfdr for each gene.

When the library sizes of two samples under comparison are clearly different, Step 0 is applied to normalize RNA-Seq data prior to implementation. The normalization procedures are described in Section . In addition, the estimated $locfdr$ reflects the probability of gene *g* being differentially expressed, and the—\citetefron-2001-empir-bayes— have shown its close connection tofalse discovery rate (FDR) controlled by Benjamini and Hochberg procedure Benjamini and Hochberg (1995). The algorithm is easy to implement, and the computation is efficient for a large G.

iDEG Algorithm for Negative Binomial Data described in Table ??

Remark 1: At Step 3, when there is no prior knowledge or strong evidence to suggest a constant dispersion across genes, the smoothing spline fit should be used. Our simulated experiments show that the smoothing spline can produce a nearly constant $\hat{\delta}_g$ in the constant dispersion case. Furthermore, the linear regression model (**??**) has slightly better performance when the dispersion is constant, but considerably worse when δ_g is not a constant across genes.

Remark 2: In most single-subject analyses, $\hat{\delta}_g$ is small. But in rare cases, when $\hat{\delta}_g \geq \frac{2}{3}$, the VST h_{nb} in Step 4 is not numerically stable. To avoid this numerical issue, we suggest replacing the VST h_{nb} by h^*_{nb} \begin{equation*}—Montgomery (2008)—, h_nbˆ*(Y_gd) = 1δgsinhˆ-1Y_gd δ_g g = 1,[?],G; d = 1,2.

Compared to h_{nb}, h^*_{nb} is less effective in stabilizing variances when μ_{gd} is small.

NUMERICAL STUDIES

Extensive numerical studies were conducted to evaluate the performance of iDEG and to compare it with existing methods, including edgeR Robinson and Smyth (2007), DEGSeq Wang *et al.* (2009), and DESeq Anders and Huber (2010), under three experimental settings:

(1) RNA-Seq data follow the Poisson distribution;

(2) RNA-Seq data follow the NB distribution, and the dispersion parameter is a constant; and

(3) RNA-Seq data follow the NB distribution, with a varying dispersion parameter δ_g.

Under each setting, single-subject RNA-Seq datasets are simulated with different percentages of DEGs, including $p = 5\%, 10\%, 15\%, 20\%$. Each experiment is repeated 1000 times, and for each time, a baseline transcriptome and a case transcriptome are generated to compose a RNA-Seq dataset. Performance of the methods are assessed by their precision, false positive rate (FPR), recall, and F_1 score, which is a harmonic mean of precision and recall. The average number of identified DEGs are also reported.

Comparison of Different Methods for the two numerical Studies.

Note: the numbers in parentheses represent the standard deviations

Study 2
proportion
Method
Precision
Recall/TPR
FPR
F1
Predicted DEG
5%
iDEG
0.957 (1.0×10^{-2})
0.733 (1.9×10^{-2})
0.002 (4.7×10^{-4})
0.83 (1.1×10^{-2})
766 (26)
edgeR
0.532 (1.1×10^{-2})
0.935 (7.7×10^{-3})
0.043 (1.9×10^{-3})
0.678 (9.0×10^{-3})
1760 (39)
DESeq
1 (0)
0.07 (3.6×10^{-2})
0 (0)
0.131 (6.1×10^{-2})

Figure 2: Comparison of F_1 scores for iDEG and other existing methods. Each point represents the average F_1 scores resulted from 1000 repeated experiments, and the horizontal bars represent one standard deviation. Each panel presents one distributional assumption of RNA-Seq data (right: Study 1 = Poisson distribution, Study 2 = NB distribution with a constant dispersion parameter, Study 3 = NB distribution with a varying dispersion parameter as a function of expression mean). FDR Benjamini and Hochberg (1995) cutoff is set to 0.1 for edgeR and DESeq; local fdr cutoff is set to 0.2 for iDEG. DESeq produced no results for the Poisson case and for some datasets in the other two panels. For visualization clarity, horizontal axes are set to different ranges.

70.35 (36)
DEGseq
0.102 (9.0×10^{-4})
0.985 (3.9×10^{-3})
0.459 (4.4×10^{-3})
0.184 (1.5×10^{-3})
9699 (85)
10%
iDEG
0.966 (8.2×10^{-3})
0.78 (1.9×10^{-2})
0.003 (8.2×10^{-4})
0.863 (9.7×10^{-3})
1616 (50)
edgeR
0.639 (8.8×10^{-3})
0.947 (5.2×10^{-3})
0.06 (2.3×10^{-3})
0.763 (6.8×10^{-3})
2966 (42)
DESeq
NA (NA)
0 (0)
0 (0)
NA (NA)
0 (0)
DEGseq
0.19 (1.6×10^{-3})
0.986 (2.8×10^{-3})
0.468 (4.5×10^{-3})
0.318 (2.3×10^{-3})
10394 (80)
15%
iDEG
0.969 (5.1×10^{-3})
0.814 (1.5×10^{-2})
0.005 (8.3×10^{-4})
0.884 (7.7×10^{-3})
2519 (54)
edgeR
0.699 (7.2×10^{-3})
0.954 (4.1×10^{-3})
0.073 (2.5×10^{-3})
0.807 (5.2×10^{-3})
4098 (44)
DESeq
NA (NA)
0 (0)
0 (0)
NA (NA)
0 (0)
DEGseq
0.266 (2.1×10^{-3})
0.987 (2.1×10^{-3})

0.48 (5.0×10^{-3})
0.419 (2.6×10^{-3})
11128 (86)
20%
iDEG
0.974 (4.2×10^{-3})
0.828 (1.5×10^{-2})
0.006 (1.0×10^{-3})
0.895 (7.8×10^{-3})
3402 (74)
edgeR
0.741 (6.0×10^{-3})
0.96 (3.2×10^{-3})
0.084 (2.6×10^{-3})
0.836 (4.1×10^{-3})
5182 (45)
DESeq
NA (NA)
0 (0)
0 (0)
NA (NA)
0 (0)
DEGseq
0.333 (2.3×10^{-3})
0.987 (1.9×10^{-3})
0.494 (5.0×10^{-3})
0.498 (2.6×10^{-3})
11858 (80)
Study 3
proportion
Method
Precision
Recall/TPR
FPR
F1
Predicted DEG
5%
iDEG
0.926 (1.5×10^{-2})
0.652 (2.2×10^{-2})
0.003 (6.3×10^{-4})
0.765 (1.4×10^{-2})
704 (29)
edgeR
0.305 (6.3×10^{-3})
0.956 (6.0×10^{-3})
0.115 (3.4×10^{-3})
0.463 (7.3×10^{-3})
3133 (65)
DESeq
0.999 (2.1×10^{-3})
0.152 (3.8×10^{-2})
0 (1.8e-05)
0.262 (5.8×10^{-2})
152 (38)
DEGseq
0.086 (6.7×10^{-4})
0.985 (3.9×10^{-3})
0.549 (3.9×10^{-3})
0.159 (1.2×10^{-3})
11409 (74)
10%
iDEG
0.945 (1.1×10^{-2})
0.708 (2.2×10^{-2})
0.005 (1.1×10^{-3})
0.809 (1.2×10^{-2})
1500 (59)
edgeR
0.447 (6.2×10^{-3})
0.96 (4.3×10^{-3})
0.132 (3.3×10^{-3})
0.61 (6.0×10^{-3})
4296 (60)
DESeq
1 (0)
0 (5.2×10^{-4})
0 (0)
0.002 (1.4×10^{-3})
1 (1)
DEGseq
0.165 (1.1×10^{-3})
0.986 (2.5×10^{-3})
0.556 (4.2×10^{-3})
0.282 (1.6×10^{-3})
11975 (76)
15%
iDEG
0.953 (7.0×10^{-3})
0.746 (1.6×10^{-2})
0.006 (1.1×10^{-3})
0.837 (9.1×10^{-3})
2349 (58)
edgeR
0.537 (5.7×10^{-3})
0.964 (3.7×10^{-3})
0.147 (3.4×10^{-3})
0.69 (4.8×10^{-3})
5384 (59)
DESeq
1 (NA)
0 (3.3e-05)
0 (0)
0.001 (NA)
0 (0)
DEGseq
0.235 (1.4×10^{-3})
0.986 (2.1×10^{-3})
0.565 (4.2×10^{-3})
0.38 (1.9×10^{-3})
12562 (73)
20%
iDEG
0.962 (4.6×10^{-3})
0.763 (1.3×10^{-2})
0.008 (1.0×10^{-3})
0.851 (7.8×10^{-3})
3175 (64)

edgeR
0.602 (5.7×10^{-3})
0.966 (2.8×10^{-3})
0.16 (3.9×10^{-3})
0.742 (4.4×10^{-3})
6419 (64)
DESeq
NA (NA)
0 (0)
0 (0)
NA (NA)
0 (0)
DEGseq
0.299 (1.6×10^{-3})
0.986 (2.0×10^{-3})
0.577 (4.2×10^{-3})
0.459 (1.9×10^{-3})
13180 (68)

Negative Binomial Distribution (NB) with a Varying Dispersion Parameter

This study assumes that $Y_{g1} \sim NB(\mu_{g1}, \delta_g)$ and $Y_{g2} \sim NB(\mu_{g2}, \delta_g)$, where δ_g is a constant across all genes. Besides these two assumptions, the data is generated by following the same procedure used in Section **??**. For the dispersion parameter, we set $\delta_g = 0.02$ for all $g = 1, \ldots, 20000$.

The middle panel of Figure **??** compares the F_1 scores for all methods. It is clear that iDEG is the best across the entire range of p, followed by edgeR and DEGseq. Since one main assumption in edgeR is constant dispersion, this setting actually favors edgeR. Nonetheless, iDEG still produces the higher F_1 scores across p compared to edgeR. When implementing edgeR, different values for the parameter BCV were tried and 0.1 was found to work the best. Therefore, 0.1 will be set as the default parameter value for the rest of this study. Although DESeq is able to identify some DEGs when p is small, its performance degrades quickly when p increases. This is partially due to DESeq treating two samples as replicates, which is improper when larger portions of DEGs are present in the transcriptome. It is observed that the average F_1 scores of all the methods are lower than those from Study 1, which may be due to the higher variation associated with the NB distribution.

Study 2 of Table suggests that iDEG works competitively in terms of having a combined high precision and low FPR among all methods. Take $p = 5\%$ as an example. Despite, its lower recall, iDEG has a substantially higher precision (0.957) and lower FPR (0.002) than edgeR (precision = 0.532; FPR = 0.043) and DEGseq (precision = 0.102; FRP = 0.459). DESeq occasionally yields high precision, however, its low recall leads to an overall consistently poor performance.

In this simulation , the RNA-Seq data is assumed to follow the NB distribution, where the dispersion parameter δg\delta_{g}δg is a function of μg1\mu_{g1}μg1. The simulation procedure is the same as the one described in Section 4.2 except that the dispersion parameter has been adapted to the one used by \citeauthoranders-2010-differ-expres (\citeyearanders-2010-differ-expres) and set δg=0.005+9/(μg1+100)\delta_{g}=0.005+9/(\mu_{g1}+100)δg=0.005+9/(μg1+100). The bottom panel in Figure ??? suggests that iDEG produces the highest F1F_{1}F1 scores across ppp. Study 3 in Table ??? has a similar pattern as Study 2 in Table ???, suggesting that iDEG has the best overall performance in terms of high precision and low FPR,regardless of whether δg\delta_{g}δg is a constant or a function of expression mean μg\mu_{g}μg.

Unequal Library Sizes

Three numerical studies (Sections **??**-) with single-subject, single-sample RNA-Seq data with unequal library sizes (where the library size of one transcriptome is 1.5 times that of the other transcriptome) were also conducted. The results are shown in the Web Figure 2. These results demonstrate that the iDEG can adjust unequal library sizes well and its performance is still superior to existing methods.

SENSITIVITY ANALYSIS

The proposed iDEG procedure makes two assumptions about the data: 1) a functional mean-variance relationship in RNA-Seq data exists, and 2) the majority of the genes are null genes. Both assumptions are commonly accepted and used in the literature; however, the performance of iDEG is unknown when these assumptions are violated. Therefore, this section examines the sensitivity of iDEG to these two assumptions.

Robustness of iDEG to Random Dispersions

We simulate RNA-Seq data from the NB distribution with δ_g drawn from a uniform distribution Uniform$(0.001, 0.1)$. In this setup, δ_g is no longer a function of μ_g. As shown in Panel A of Figure **??**, all methods perform worse than in previous studies, but iDEG is still best among the four in terms of the highest F_1 scores.

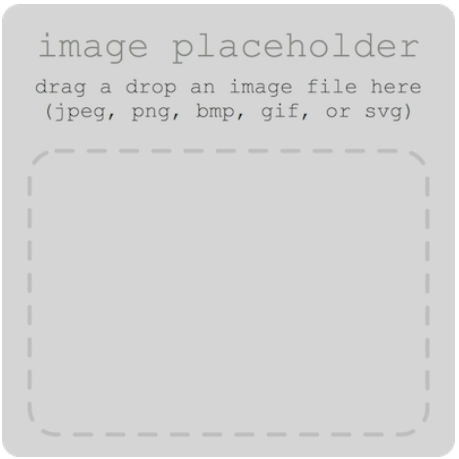

Figure 3: iDEG is robust to its the assumptions. Panel A indicates robustness of iDEG to the assumption that δ_g is a function of expression mean μ_g. The F_1 scores of iDEG are the highest among the four methods, when the values of δ_g are randomly drawn from a uniform distribution Uniform$(0.001, 0.1)$. Panel B indicates robustness of iDEG that the majority of the genes are null genes. In this panel, the F_1 scores of edgeR approach the scores of iDEG at unrealistically high percentages of DEGs.

Robustness of iDEG to high percentages of DEG

The four methods were compared on RNA-Seq data with p as high as 40%. As shown in Panel B of Figure **??**, iDEG still performs better than edgeR even though the latter does not make the low-p assumption. Note, $p = 40\%$ is an unrealistic extreme case in biology since this would result in nearly half of the genes in the transcriptome being associated or altered by a disease.

APPLICATION OF IDEG TO TRIPLE NEGATIVE BREAST CANCER (TNBC) STUDY

The method iDEG was applied to a triple negative breast cancer (TNBC) dataset queried from TCGA, which has been described earlier in Section . Recall this single-subject RNA-Seq data provide measures of a breast tumor transcriptome and a surrounding healthy tissue transcriptome of a TNBC African American patient (Patient ID: TCGA-GI-A2C9). The goal of this study is to apply iDEG for the discovery of individualized DEGs for this single patient.

The expression counts are assumed to follow the NB distribution and were normalized as described in Section . Since there is no prior evidence suggesting the constant dispersion across all genes, a smoothing spline (Section **??**) was fit to estimate the relationship between δ_g and μ_g. The empirical null obtained by iDEG is $N(0.065, 0.837^2)$. A local false discovery rate (*fdr*) was produced for each of the 20,501 genes. Figure **??** indicates that DEGs are determined with an adaptive cutoff that accounts for the high noise of the lowly expressed genes. For patient TCGA-GI-A2C9, iDEG identified 1,430 DEGs (approximately 7% of all genes) by controlling a local fdr below 20%. Table **??** displays the top 10 DEGs detected by iDEG, in the ascending order of a local fdr. In contrast, edgeR identified 9,921 genes as DEGs (FDR $\leq$ 0.1), which amounts to almost half of the transcriptome. DESeq, on the other hand, only identified 194 genes (FDR $\leq$ 0.1), which is far fewer than one would expect from a cancer patient. While it is impossible to know which genes are truly differentially expressed, the range of the number of DEGs in cancer patients is common knowledge for cancer researchers Cancer Genome Atlas Network (2012). We conclude that iDEG can identify a reasonable number of DEGs for this patient.

Let us now take a careful look at the genes listed in Table **??**. Adiponectin (gene product of gene ADIPOQ) and leptin (gene product of LEP) are considered mediators for the association of breast cancer with obesity, a major risk factor for breast cancer Grossmann (2010). It has been shown that the reduction in adiponectin and leptin levels increases breast cancer risk Miyoshi and et al (5699); Duggan and et al (2011); Karim and et al (2016), and the treatment of adiponectin induces growth arrest and apoptosis of breast cancer cell lines Jarde and et al (1197); Kang (1263). Our finding of the decreased expression of ADIPOQ and LEP suggests that obesity may substantially contribute to this patient's cancer development. On the other hand, although PLA2G2A has not been extensively studied in breast cancer, many studies have shown that it inhibits invasion and metastasis of gastric and colon cancer Ganesan and et al (4277) and may predict survival Xing (2011). Informed by her individualized DEG, we speculate that successful treatments for gastric and colon cancer may benefit patient TCGA-GI-A2C9. \citeauthorbubnov-2012-hypermethylation (\citeyearbubnov-2012-hypermethylation) has demonstrated the down-regulation of TUSC5 induced by DNA methylation in breast cancer. In contrast to mutated genes, DNA methylation is reversible. If the TUSC5 of patient TCGA-GI-A2C9 is suppressed by DNA methylation, pharmacologic inhibition of methylation-mediated TUSC5 suppression could potentially treat this patient Baylin (2005). Futher investigation of these discovered DEGs may unveil this patient's disease etiology, progression, and possible therapeutic targets, which can eventually lead to an improved personalized treatment plan.

The top-10 hits, smallest local false discovery rate (*fdr*), DEGs identified by iDEG and their local fdr for TNBC patient TCGA-GI-A2C9.

Gene	*fdr*	Z
ADIPOQ	2.85e-34	-11.17
PLA2G2A	2.85e-34	-11.65
PI16	1.15e-33	-10.78
LEP	2.25e-33	-10.70
SFTPB	1.44e-32	-10.59
IL33	4.24e-31	-10.36
TUSC5	6.74e-31	-10.32
CSF3	2.89e-29	-10.04
COL6A6	3.24e-29	-10.04
CCL21	1.99e-28	-9.90

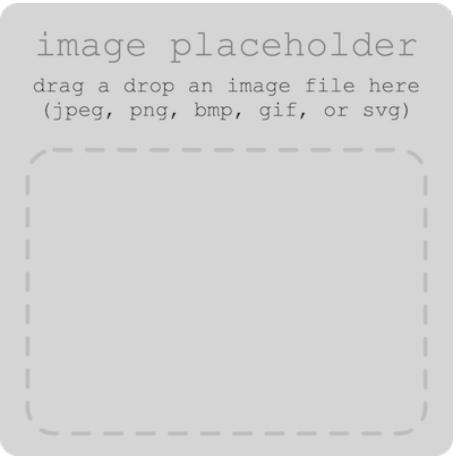

Figure 4: Adaptive cutoffs were determined by iDEG. In terms of the log_2 transformed fold change, iDEG, in general, sets cutoffs with larger values for lowly expressed genes and cutoffs with lower values for the highly expressed genes, which accounts for the precision of the gene expression measurement. Data are displayed in a typical MA plot. The x axis is the average of the log_2 transformed expression counts, and the y axis is the log_2 transformed fold change.

DISCUSSION

By focusing on one patient at a time in which each subject serves as his/her own control, single-subject analyses, including the one we propose, have the potential to ascertain meaningful biomolecular mechanisms for decision-making in precision medicine Gardeux *et al.* (1116). However, the prohibitive cost and access to clinical tissue in a single subject undermines the replication requirements of conventional statistical methods. In this work, we introduce a novel and powerful method for identifying DEGs based on only two transcriptomes for a single subject (case vs. baseline transcriptome). The core idea is the application of variance-stabilizing transformation (VST), which effectively solves the single-subject, single-sample problem and makes it possible to "borrow strength" across genes. Through simulation studies and a clinical dataset analysis, it was demonstrated that iDEG has a high accuracy of discovery even when gene expression counts are over-dispersed.

While the simulations demonstrate that iDEG presents increased accuracy at both precision (positive predictive value) and recall (sensitivity) over other methods, there are some caveats and potential extensions. First, iDEG strives to mine the most information from limited data; however, we need to keep in mind that no statistical inferences can replace data Hansen *et al.* (2011), and that replication is still preferable if the tissue is available and the associated cost is reasonable. Second, the application of iDEG is not restricted to RNA-Seq data but also applicable to count data in general, such as immunoprecipitated DNA Ross-Innes (2012) (e.g., ChIP-Seq), proteomic spectral counts Johnson (4351), protein antibody arrays, or metagenomics data, that follow Poisson or Negative Binomial distribution with the parallel structure. An important extension of iDEG can be made by incorporating suitable variance-stabilizing techniques that are suitable for high-throughput data following other distributions. Another valuable extension would be the incorporation of external knowledge, such as a gene ontology, to define a set of genes and aggregate gene-level metrics to a gene set Li *et al.* (2017a); Schissler *et al.* (2015); Li *et al.* (2017b). Lastly, future single-subject experiments may study more than two conditions beyond the current Case-vs.-Baseline design; therefore, it would be interesting to extend iDEG to identify DEGs under multiple conditions or multiple 'omics measures.

SUPPLEMENTARY MATERIALS

Web Appendices and Figures referenced in Sections , , and are available with this paper at the Biometrics website on Wiley Online Library

REFERENCES

Anders, S. and Huber, W. (2010). Differential Expression Analysis for Sequence Count Data. *Genome biology* **11,** R106.

Anscombe, F. J. (1948). The Transformation of Poisson, Binomial and Negative-Binomial Data. *Biometrika* **35,** 246.

Baylin, S. B. (2005). Dna Methylation and Gene Silencing in Cancer. *Nature Clinical Practice Oncology* **2,** S4–S11.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the royal statistical society. Series B (Methodological)* pages 289–300.

Cancer Genome Atlas Network (2012). Comprehensive Molecular Portraits of Human Breast Tumours. *Nature* **490,** 61–70.

Dietze, E. C., e. a. (2015). Triple-Negative Breast Cancer in African-American Women: Disparities Versus Biology. *Nature Reviews Cancer* **15,** 248–254.

Duggan, C. and et al (2011). Associations of Insulin Resistance and Adiponectin with Mortality in Women with Breast Cancer. *Journal of Clinical Oncology* **29,** 32–39.

Ganesan, K. and et al (4277). Inhibition of Gastric Cancer Invasion and Metastasis by Pla2g2a, a Novel -Catenin/tcf Target Gene. *Cancer Research* **68,** 4277–4286.

Gardeux, V., Berghout, J., Achour, I., Schissler, A. G., L. Q., Kenost, C. Li, J., Shang, Y., Bosco, A., Saner, D., and et al (1116). A Genome-By-Environment Interaction Classifier for Precision Medicine: Personal Transcriptome Response to Rhinovirus Identifies Children Prone to Asthma Exacerbations. *Journal of the American Medical Informatics Association* **24,** 1116–1126.

Grossmann, M. E., e. a. (2010). Obesity and Breast Cancer: Status of Leptin and Adiponectin in Pathological Processes. *Cancer and Metastasis Reviews* **29,** 641–653.

Hansen, K. D., W. Z., Irizarry, R. A., and Leek, J. T. (2011). Sequencing Technology Does Not Eliminate Biological Variability. *Nature Biotechnology* **29,** 572–573.

Jarde, T. and et al (1197). Involvement of Adiponectin and Leptin in Breast Cancer: Clinical and in Vitro Studies. *Endocrine Related Cancer* **16,** 1197–1210.

Johnson, E. K., e. a. (4351). Proteomic Analysis Reveals New Cardiac-Specific Dystrophin-Associated Proteins. *PLoS ONE* **7,** e43515.

Kaiser, J. (2015). Obama Gives East Room Rollout to Precision Medicine Initiative. *Science* .

Kang, J. H., e. a. (1263). Adiponectin Induces Growth Arrest and Apoptosis of Mda-Mb-231 Breast Cancer Cell. *Archives of Pharmacal Research* **28,** 1263–1269.

Karim, S. and et al (2016). Low Expression of Leptin and Its Association with Breast Cancer: A Transcriptomic Study. *Oncology reports* **36,** 43–48.

Li, Q., Schissler, A. G., G. V., Berghout, J., Achour, I., Kenost, C. Li, H., Zhang, H. H., and Lussier, Y. A. (2017a). Kmen : Analyzing Noisy and Bidirectional Transcriptional Pathway Responses in Single Subjects. *Journal of biomedical informatics* **66,** 32–41.

Li, Q., Schissler, A. G., G. V., Achour, I., Kenost, C., Berghout, J. Li, H., Zhang, H. H., and Lussier, Y. A. (2017b). N-Of-1-Pathways Mixenrich : Advancing Precision Medicine Via Single-Subject Analysis in Discovering Dynamic Changes of Transcriptomes. *BMC Medical Genomics* **10,** 27.

Miyoshi, Y. and et al (5699). Association of Serum Adiponectin Levels with Breast Cancer Risk. *Clinical Cancer Research* **9,** 5699–5704.

Montgomery, D. C. (2008). Em Design and Analysis of Experiments. John Wiley & Sons.

Mortazavi, A. and et al (2008). Mapping and Quantifying Mammalian Transcriptomes by Rna-Seq. *Nature Methods* **5,** 621–628.

Robinson, M. D. and Smyth, G. K. b. (2007). Small-Sample Estimation of Negative Binomial Dispersion, with Applications to Sage Data. *Biostatistics* **9,** 321–332.

Ross-Innes, C. S., e. a. (2012). Differential Oestrogen Receptor Binding Is Associated with Clinical Outcome in Breast Cancer. *Nature* **481,** 389–393.

Schissler, A. G., G. V., Li, Q., Achour, I., Li, H., Piegorsch, W. W., and Lussier, Y. A. (2015). Dynamic Changes of Rna-Sequencing Expression for Precision Medicine N-Of-1-Pathways Mahalanobis Distance Within Pathways of Single Subjects Predicts Breast Cancer Survival. *Bioinformatics* **31,** i293–i302.

Schork, N. J. (2015). Personalized Medicine: Time for One-Person Trials. *Nature* **520,** 609–611.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J., Nobel, A. Deng, S., Johnsen, H., Pesich, R., Geisler, S., and et al (8418). Repeated Observation of Breast Tumor Subtypes in Independent Gene Expression Data Sets. *Proceedings of the National Academy of Sciences* **100,** 8418–8423.

Topol, E. J. (2014). Individualized Medicine from Prewomb to Tomb. *Cell* **157,** 241–253.

Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2009). Degseq: an R Package for Identifying Differentially Expressed Genes from Rna-Seq Data. *Bioinformatics* **26,** 136–138.

Wild, C. P. and Stewart, B. W. (2014). Em World Cancer Report 2014. World Health Organization.

Xing, X.-F., e. a. (2011). Phospholipase A2 Group Iia Expression Correlates with Prolonged Survival in Gastric Cancer. *Histopathology* **59,** 198–206.