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

Qike Li[*,1,2,4], Samir Rachid Zaim[†,1,2,4], Dillon Aberasturi[1,2,4], Joanne Berghout[1-3], Haiquan Li[1,2,4], Colleen Kenost[1,2,5], Helen Hao Zhang[4,7*], Yves A. Lussier[1-6*]

Center for Biomedical Informatics and Biostatistics(CB2)[1], Department of Medicine[2], The Center for Applied Genetics and Genomic Medicine[3], Graduate Interdisciplinary Program in Statistics[4], BIO5 Institute[5], University of Arizona Cancer Center[6], Department of Mathematics[7], The University of Arizona, Tucson, AZ 85721, USA
Email: hzhang@math.arizona.edu, daberasturi@email.arizona.edu, haiquan@email.arizona.edu, ckenost@email.arizona.edu, hzhang@math.arizona.edu, yves@email.arizona.edu

Accurate profiling of gene expression in a single subject has the potential to be a powerful precision medicine tool, useful for unveiling individual disease mechanisms and responses. However, most expression analysis tools for RNA-sequencing (RNA-Seq) data require replicate samples to estimate gene-wise data variability and make inferences, which is costly and not easily obtainable in clinical practice. We propose the iDEG method to identify individualized Differentially Expressed Genes in a single subject sampled under two conditions without replicates, i.e. a baseline sample (unaffected tissue) vs. a case sample (tumor). iDEG borrows information across different genes from the same individual using a partitioned window to strategically bypass the requirement of replicates per condition. It then transforms RNA-Seq data such that, under the null hypothesis, differences of transformed expression counts follow a distribution and variance calculated across a local partition of related transcripts at baseline expression. This transformation enables modeling 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 simulation studies demonstrate iDEG's substantially was the only technique keeping high precision (>90%), recall (>75%) and low false positive rate (<1%) accuracy under thousands of scenarios when compared to DESeq, edgeR, and DEGseq. Software available at:
http://www.lussiergroup.org/publications/iDEG

Keywords: transcriptome, mRNA expression, RNAseq, differentially expressed genes, single-subject, N-of-1, RNA-Seq, iDEG.

[†] Contributed equally

[a] This work was supported in part by The University of Arizona Health Sciences CB2, the BIO5 Institute, NIH (U01AI122275, CA023074, 1UG3OD023171, 1S10RR029030)

* Corresponding authors

Preprint of an article submitted for consideration in the Proceedings of the Pacific Symposium on Biocomputing, World Scientific Publishing Company

1. Background

Precision medicine aims to deliver "the right treatments, at the right time, to the right person"[1]. However, clinical research, medicine, and pharmacology need new tools to achieve that goal. The prevailing system of one-size-fits-all drug development has led to the ten top-grossing USA drugs being ineffective for more than 75% of users[2], and these patients typically cannot be identified until after therapeutic failure has occurred. The success of precision medicine hinges on identifying the precise aberrant mechanisms at play during an individual's disease course[3] to optimize treatment based on that individual's biology.

Single-subject RNA sequencing (**RNA-Seq**) analysis considers one patient at a time, with the goal of revealing an individual's altered transcriptomic mechanisms. Relative to traditional cohort-based analyses, a major challenge of single-subject RNA-Seq analysis is the estimation of gene expression variance which is required to identify differentially expressed genes (**DEGs**). In cohort-based methods, gene variance is calculated across a large heterogenic set of samples, and the statistical methods employed leverage and rely on those replicates. However, they also emphasize consistent and average responses which may not accurately represent a single patient when the disease is heterogenous or stratified. Alternatively, the variance can be assessed between two conditions in one subject and three replicates. Yet, obtaining sufficient isogenic replicates for one subject to answer more precision questions poses a major difficulty 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. Even though there is a great body of work for identifying DEGs in RNA-Seq data[4-8] and frameworks for N-of-1 studies either for a single analyte or by pooling gene products in pathways[9-13], very few methods have been designed or validated at the gene level to determine the effect size and statistical significance of a single-subject, single RNA-Seq studies in two conditions without replicates (**TCWR**)[14].

Three critical obstacles hinder the analysis of single-subject **TCWR** studies. These include i) *patient-level inferential capability* in absence of biological replicates, ii) sensitivity to *fold-change inflation* in low-expression genes, and iii) *rigid parametric data assumptions* for variance estimation. To bridge the knowledge gap and develop a technique that can overcome the current technical limitations in analyzing RNA-Seq data, we developed **iDEG** (**i**ndividualized **D**ifferentially **E**xpressed **G**enes). iDEG is a novel RNA-Seq analysis tool designed to analyze paired transcriptomic samples from a single individual to effectively compare case vs. baseline measurements (e.g., tumor vs. healthy tissue) in the absence of biological replicates. In order to bypass the variance estimation limitations in single-subject, single-replicate studies, iDEG "borrows" information across genes of similar baseline expression levels and applies a variance-stabilizing transformation to estimate a gene's distribution. iDEG provides predictions that are robust to fold-change inflation and to constricting data assumptions for identifying individualized DEGs in absence of biological replicates. Our simulation studies and exemplar study in a triple negative breast cancer dataset from The Cancer Genome Atlas illustrate iDEG's superior performance to existing methods as well its direct application in real datasets. This study demonstrates iDEG's utility for analyzing single-subject, single replicate RNA-Seq studies to advance precision medicine.

2. Methods

2.1. *The iDEG algorithm: iDEG*

The iDEG algorithm (**Figure 1**) is an easy-to-implement, single-command function written in R[15] with a computation speed of one second for identifying a subject's DEGs on 8GB Ram computer. The subsequent sections expand on **Figure 1**.

2.2. *Modeling read counts via a re-parameterized Negative Binomial (NB) distribution*

We model read counts Y_{gd} as following a re-parameterized negative binomial distribution with mean μ_{gd} and dispersion δ_g. Thus, $Y_{gd} \sim NB(\mu_{gd}, \delta_g)$ with the following probability mass function, mean, and variance, respectively. Since both transcriptomes are sequenced separately, they are treated as independent.

$$P(y_{gd}|\mu_{gd}\delta_g) = (1+\delta_g\mu_{gd})^{-\frac{1}{\delta_g}} \frac{\Gamma\left(y_{gd}+\frac{1}{\delta_g}\right)}{y_{gd}!\,\Gamma\left(\frac{1}{\delta_g}\right)} \left(\frac{\delta_g\mu_{gd}}{1+\delta_g\mu_{gd}}\right)^{y_{gd}} ; E(Y_{gd}) = \mu_{gd},\ Var(Y_{gd}) = \mu_{gd} + \delta_g\mu_{gd}^2 \quad \textbf{(Eq.1)}$$

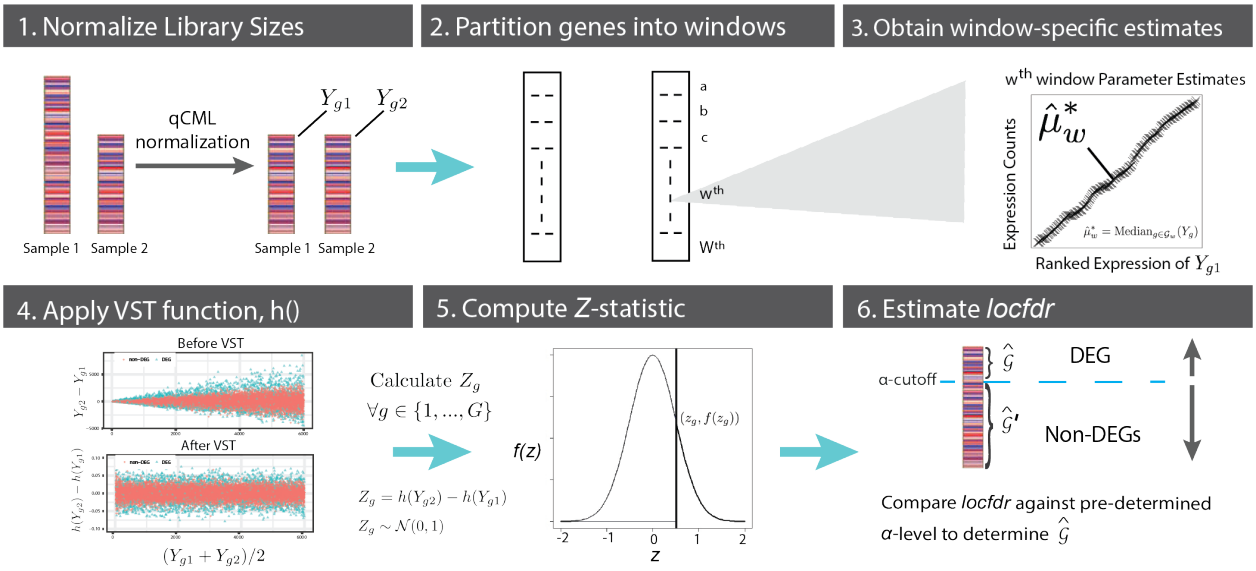

Fig. 1. The iDEG Algorithm. 1) Normalize unequal library sizes if necessary. 2) Partition transcriptome into percentile-based windows using ranked baseline expression. 3) For each window: estimate mean expression, variance, and dispersion parameters. 4) Apply the Variance Stabilizing Transformation (VST) for each gene expression count. 5) Calculate the standard normal summary statistic "Z_g" for each gene expression count "g". 6) Determine the identified DEG set "$\hat{\mathcal{G}}$" based on a pre-determined ***α*-cutoff.**

To identify DEGs from a pair of transcriptomes, we must test multiple hypotheses H_0: $\mu_{g1} = \mu_{g2}$, $g = 1, \ldots, G$, where μ_{g1} and μ_{g2} are the theoretical mean expression levels for each gene "g" in sample 1 and sample 2, respectively. We define the DEG set by $\mathcal{G} = \{g: \mu_{g1} \neq \mu_{g2}, g = 1, \ldots, G\}$ and its set-theoretic complement of non-differentially expressed genes, or "null gene set" by $\mathcal{G}' = \{1, \ldots, G\} \setminus \mathcal{G}$. If there were replicates, each hypothesis could be tested with a two-sample comparison,

using Welch's t-test statistic: $t_g = (\bar{Y}_{g1} - \bar{Y}_{g2})/[(S_{g1}^2/N_{g1} + S_{g2}^2/N_{g2})^{1/2}]$ were $\bar{Y}_{g1}$ and $\bar{Y}_{g2}$, S_{g1} and S_{g2}, and N_{g1} and N_{g2} are each groups' respective sample mean, standard deviation, and size. However, when there is only one observation for Y_{g1} and one for Y_{g2}, neither S_{g1}^2 nor S_{g2}^2 are computable. We thus propose iDEG: an algorithm that transforms Y_{g1} and Y_{g2}, such that a simple function of the transformation allows for modeling all genes with the same distribution. This is done by pooling the genes together and estimating their common variance, hence bypassing the single-subject, single-replicate limitation.

2.3. *Normalize read counts with unequal library sizes (Fig. 1, Panel 1)*

In practice, unequal DNA library sizes may exist; thus, the first step is to normalize library sizes if necessary. We use the qCML (quantile-adjusted conditional maximum likelihood) procedure by Robinson and Smyth[16], for normalization, and subsequently, iDEG is applied.

2.4. *Partition genes into windows to estimate local mean and variance (Fig. 1, Panel 2)*

Genes of comparable expression levels are assumed to behave similarly in iDEG. Thus, after normalization, the next step is to group genes into W non-overlapping windows of similar expression levels to approximate each window's local mean and variance parameters. In the re-parameterized NB distribution, the variance of a given gene, g, is a function of its mean, μ_g, and dispersion, δ_g. Thus, the genes are partitioned to obtain their local, window-specific parameters. We define the wth window by the $(w-1)^{th}$ and $(w)^{th}$ percentiles, for w=1,…,W. $\mathcal{G}_w = \{g: (w-1)^{th}\ percentile\ of\ Y_{g1} < Y_{g1} < w^{th}\ percentile\ of\ Y_{g1}\}$. To provide robust parameter estimates, we recommend a large positive integer, W, so that each window contains between 150 and 200 genes. However, the final predictions are not overtly sensitive to the choice of W (< 10% difference, data not shown – available upon request).

2.5. *Compute each window's parameters (Fig. 1, Panel 3)*

As seen in (Eq. 1), $Var(Y_{gd}) = \mu_{gd} + \delta_g \mu_{gd}^2$. Therefore, when estimating variance locally, we are actually estimating the dispersion parameter δ_g for each gene count. This value is required for the variance-stabilizing transformation (VST) h in equation (**Eq. 3; Fig. 1 Panel 4**). However, variance cannot be estimated with only a single observation. For RNA-Seq data analysis, one common assumption is that the dispersion δ_g is equal across samples 1 and 2, and that dispersion is a function q of the mean, μ_g[17-19]. Thus, in the absence of replicates, we partition genes into small windows to estimate **the functional mean-dispersion relationship,** $\delta_g = q(\mu_g)$, and hence the variance. We propose a two-step nonparametric procedure to obtain: (i) an initial estimate of δ_g by pooling genes locally; and (ii) a refined estimate of δ_g by estimating $q(\mu_g)$ with a smooth curve-fitting technique. In this approach, all non-differentially expressed (null) genes belonging to the same window $\mathcal{G}_w$ roughly have the same mean μ_w^* and the dispersion value δ_w^*. Thus, $Y_{gd} \sim NB\ (\mu_w^*, \delta_w^*),\ \forall g \in \mathcal{G}_w,\ d = 1,2$ where μ_w^* and δ_w^* are the window-specific mean and dispersion values for null genes in $\mathcal{G}_w$, while d specifies if the count comes from sample 1 or 2.

The initial window estimates μ_w^* and σ_w^{2*} are obtained by $\hat{\mu}_w^* = Median\left(Y_{g1}\right)\ \forall g \in \mathcal{G}_w$ and $\hat{\sigma}_w^* = Median\left(\left|Y_{g1} - Median\left(Y_{g1}\right)\right|\right)\ \forall g \in \mathcal{G}_w$, respectively. Since $Var\left(Y_{gd}\right) = \sigma_{gd}^2 = \mu_{gd} + \delta_g \mu_{gd}^2$, we estimate δ_w^* with $\hat{\delta}_w^* = (\sigma_w^{2*} - \hat{\mu}_w^*)/\hat{\mu}_w^{2*}$, for all windows $w = 1, \cdots, W$. To further improve the estimate of the dispersion parameter, δ_g, a smoothing spline technique is used to fit a functional mean-dispersion relationship, $\delta_g = q(\mu_g)$, by solving the following optimization problem

$$\hat{q}_\lambda = \min_{q \in Q} \sum_{w=1}^{W} \left\{\hat{\delta}_w^* - q\left(\hat{\mu}^*{}_w\right)\right\}^2 + \lambda \int [q''(t)]^2\, dt, \qquad \textbf{(Eq. 2)}$$

where Q is the second-order Sobolev space on [0, 1] containing q, and λ is a smoothing parameter (selected via generalized cross validation)[20]. After the fitted curve $\hat{q}_\lambda$ is obtained in (Eq. 2), the refined estimate of $\hat{\delta}_g$ is computed as $\hat{\delta}_g = \hat{q}_\lambda\left(\hat{\mu}_w^*\right), \quad \forall g \in \mathcal{G}_w, \quad w = 1, \cdots, W$.

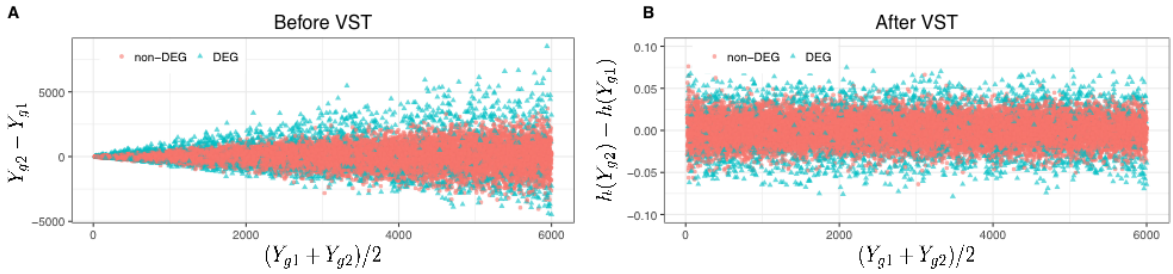

Fig. 2. Variance Stabilizing Transformation (VST). Panel A depicts the raw difference $\boldsymbol{D_g = Y_{g1} - Y_{g2}}$ for 20,000 simulated genes (simulation **Methods 2.9**), suggesting that the variance of $\boldsymbol{D_g}$ increases as the mean $\boldsymbol{\mu_{g1}}$ increases; hence, there is no uniform cutoff to differentiate DEGs and null genes. Panel B illustrates that, for null genes, VST makes the variance of $\boldsymbol{D_g^* = h(Y_{g1}) - h(Y_{g2})}$ constant regardless of mean $\boldsymbol{\mu_{g1}}$.

2.6. *Apply the Variance Stabilizing Transformation ($h(Y_{gd})$) to each gene (Fig.1, Panel 4)*

After fitting δ_g, we apply the variance stabilizing transformation h to the counts, $Y_{gd}\ \forall g \in \mathcal{G}_w$:

$$h\left(Y_{gd}\right) = \frac{1}{\sqrt{\delta_g}} \sinh^{-1} \sqrt{Y_{gd}\delta_g} + \left(\sqrt{\frac{1}{\delta_g} - 1}\right) \sinh^{-1} \sqrt{\frac{Y_{gd} + \frac{3}{4}}{\frac{1}{\delta_g} - \frac{3}{2}}}, \forall g \in \mathcal{G}_w \ \textbf{(Eq. 3)}.$$

This transformation[21] of the expression Y_{gd} in each window w, results in an approximately constant variance across all windows of the transcriptome (**Fig. 2**), regardless of the expression mean, μ_{gd}. That is $h(Y_{gd}) \dot{\sim} N\left(h\left(\mu_{gd}\right), 1\right)$, where d specifies if the count comes from sample 1 or 2. Therefore, the difference of the two independent normal random variables (e.g., $h\left(Y_{g1}\right) - h(Y_{g2})$) approximately follows a common normal distribution with mean 0 and a constant variance: $Z_g = (1/\sqrt{2})\left[h\left(Y_{g1}\right) - h\left(Y_{g2}\right)\right] \sim N(0,1)$. We suggest replacing $\sqrt{2}$ by a robust estimate of standard deviation (e.g., median absolute deviation)[22]. In most single-subject analyses, the estimated dispersion parameter, $\hat{\delta}_g$, is small, but when $\hat{\delta}_g \geq 2/3$, the VST $h(Y_{gd})$ is not numerically stable. To avoid this issue, we suggest replacing h with h^* [23], where $h^*\left(Y_{gd}\right) = (1/\sqrt{\delta_g}) \sinh^{-1} \sqrt{Y_{gd}\delta_g}, \quad g = 1, \cdots, G; \quad d = 1,2$. If a negative value of $\hat{\delta}_0$ is obtained, we conservatively set it to zero to assume a larger variance.

2.7. *Compute the summary statistic for each gene (Fig.1, Panel 5)*

In the context of noisy data and large-scale inference, performing individual tests neglects the parallel structure of RNA-Seq data. Moreover, actual data mean and variance may not be close to their theoretical values of 0 and 1 due to various reasons (e.g., correlation across genes, correlation between samples, or failed mathematical assumptions)[24]. Therefore, we estimate an empirical null distribution $N(\mu_0, \sigma_0)$ to test these individual hypotheses.

Since differentially and non-differentially expressed genes generally follow different distributions, the probability density function of $Z_g, f(z)$, is naturally modeled by a two-group mixture: $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$. Here, f_0 and f_1 are the probability density functions of genes in $\mathcal{G}'$ and in $\mathcal{G}$, while π_0 and $\pi_1 = 1 - \pi_0$ are their respective membership proportions. We assume a normal distribution following previous work from Dean and Raftery[25] that applied a two-group mixture model to identify differentially expressed genes, assuming a normal distribution for the null genes and a uniform distribution for the DEGs. However, we relax their assumptions for the marginal distribution and assume an exponential family. We approximate $f(z)$ using a smooth K-parameter exponential family distribution, $f(z) = \exp\{\sum_{k=0}^{K} \beta_k z^k\}$, and estimate the parameters $(\beta_0, \beta_1, \cdots, \beta_K)^T$ using Efron's approach[26].

2.8. *Estimate the local false discovery (locfdr) for each and identify DEGs (Fig. 1, Panel 6)*

Finally, to control the false discovery rate (fdr), we adopt Efron's idea[27-29] to estimate the local fdr (locfdr) using the R package *locfdr* and estimate π_0, $f_0(z)$ by maximum likelihood. Efron et al.[30] have shown locfdr's close connection to the BH false discovery rate procedure[31]; therefore, after estimating $locfdr(z_g)$, it identifies differentially expressed genes by comparing $locfdr(z_g)$ to a pre-specified α-cutoff value. The identified set of differentially expressed genes by the iDEG procedure is denoted by $\widehat{\mathcal{G}}$.

2.9. *Simulations (Fig. 2-3)*

To compare the performance of iDEG to existing methods – including edgeR[16], DEGSeq[8], DESeq[18], and NOISeq[6] – extensive numerical studies were conducted assuming that RNA-Seq data follow the NB distribution with a varying dispersion parameter δ_g. Of note, all these methods assume the negative binomial distribution for data, which is used in the simulation; with the exception of NOISeq that is nonparametric and DEGseq which assumes a binomial distribution. Baseline (normal tissue) and case (tumor sample) transcriptomes are both simulated and assumed to contain G = 20,000 genes; the library size of one transcriptome is 1.5 times larger than the other one. The single-subject RNA-Seq datasets are simulated with different percentages of DEGs, including DEG percentage = 5%, 10%, 15%, 20%, 25%, 30%, 35%, and 40%. Each experiment is repeated 1000 times. $Y_{g1} \sim \mathrm{NB}(\mu_{g1}, \delta_g)$ and $Y_{g2} \sim \mathrm{NB}(\mu_{g2}, \delta_g)$, where the dispersion parameter δ_g has been set to $\delta_g = 0.005 + 9/(\mu_{g1} + 100)$, per Anders and Huber[18]. Gene expression means, μ_{g1}, are sampled from: $P(\mu_{g1}) = (1/500)e^{-(1/500)\times \mu_{g1}}, \quad g = 1, \cdots, 20000$. For the case transcriptome, we set $\mu_{g2} = \mu_{g1}$ for $g \in \mathcal{G}'$ and $\mu_{g2} = d^s \mu_{g1}$ for $g \in \mathcal{G}$, where $s = (-1)^b$ and $b \sim$ Bernoulli(0.5) is a random variable. Here, s indicates increasing expression (s = 1) or decreasing expression (s =

-1) of a gene in the case transcriptome relative to baseline. Finally, for each gene g, we simulate one observation for Y_{g1} and Y_{g2} respectively, and test the hypothesis $\mu_{g1} = \mu_{g2}$. At each iteration, a baseline and a case transcriptome are generated to simulate a distinct RNA-Seq dataset. Methods are assessed by their Precision, Recall, and FPR, and F_1 score, $F_1 = (2 \cdot (\text{precision} * \text{recall})/(\text{precision} + \text{recall})$. The average number of identified DEGs are also reported.

Of note, we excluded GFOLD[7] from the comparison, as it only ranks genes without providing a measure of significance. In clinical practice, selecting a rank cutoff for GFOLD would be arbitrary and less useful than a method providing significance.

3. Results

Table 1 and **Figure 3** depict the accuracies obtained in the simulation.

Table 1. At 10% FDR, iDEG obtains the best precision and F_1 score in most single-subject TCWR simulations. Of note, edgeR, DEGseq, and DESeq were not designed nor validated for studies without replicates; however, the authors proposed some parameters to utilize them in these conditions. NOISeq-sim is a NOISeq solution for studies without replicates and obtains a better F1 score when 8,000 genes are dysregulated among 20,000 **(Fig. 3).**

Proportions of DEGs seeded	Method	Precision	Recall (TPR)	FP	*Predicted DEGs*
	iDEG	**0.93**	**0.65**	**57**	***700***
	edgeR	0.31	0.96	2,090	*3,119*
1,000 out of 20,000	NOISeq-sim	0.412	0.89	1140	*2,163*
	DESeq	1.0	0.16	0	*162*
	DEGseq	0.086	0.99	10,450	*11,397*
	iDEG	**0.97**	**0.76**	**112**	***3,136***
	edgeR	0.60	0.97	2,560	*6,405*
4,000 out of 20,000	NOISeq-sim	0.747	0.91	1120	*4,897*
	DESeq	Not applicable	0	0	*0*
	DEGseq	0.3	0.99	9,280	*13,160*
	iDEG	**0.98**	**0.80**	**120**	***6,484***
	edgeR	0.77	0.97	2,400	*10,145*
8,000 out of 20,000	NOISeq-sim	0.893	0.91	840	*8,177*
	DESeq	Not applicable	0	0	*0*
	DEGseq	0.51	0.99	7,560	*15,455*

4. Discussion and Limitations

Low expression levels are extremely susceptible to unstable fold-change estimation, as a 5-fold increase from 2 to 10 counts on a dynamic range of 0 to 100,000 should not be treated equivalently to that between 10,000 and 50,000. Standard practice filters out genes with counts below a certain threshold (typically 5 or 10). However, this solution does not address fold change inflation above the threshold (e.g., FC>2 at 15 counts), nor how to compare distinct FCs at different expression levels. Alternatively, favoring absolute count difference to identify DEGs leads to a systemic bias towards genes with high expression. Conversely, favoring FC results in a systemic bias towards lowly expressed genes. Either of these solutions yields higher false positive rates. Thus, iDEG addresses the dilemma of dealing with fold change inflation by comparing FC values relative to their expression levels, while conventional approaches impose stringent data assumptions that may

compromise downstream inferential processes. These restrictions occur either by methods always assuming rather than estimating a distribution and its parameters or by treating case and control observations as replicates to compute the variance in their absence.

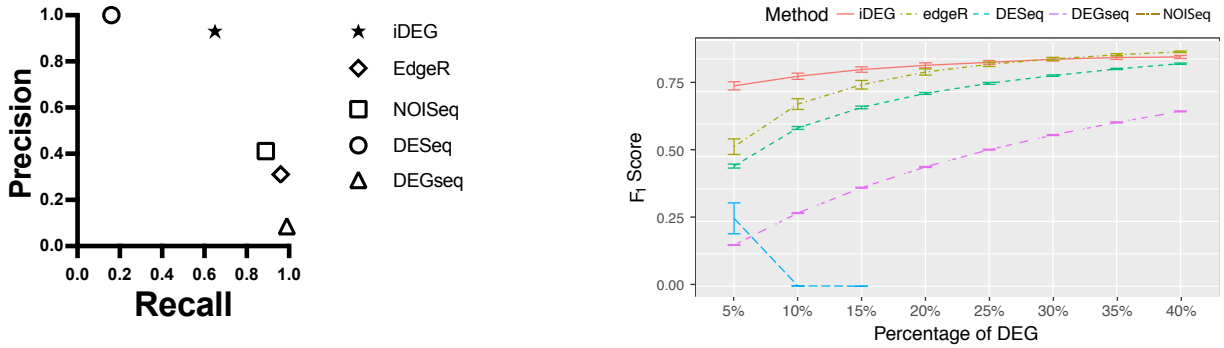

Fig. 3. At 10% FDR, iDEG's accuracies lead to superior clinical utility for precision medicine. At FDR<10%, NOIseq, DESeq, and EdgeR operate at high recall and precision <45%, while DEGseq operates at high precision and low recall (<20%). Only iDEG provides a balanced precision-recall in clinical conditions where one cannot change the stated FDR of a method because we generally only obtain one sample in each conditions for one patient. **Left panel**. Precision recall curve at 10% FDR for 1,000 seeded DEGs among 20,000 transcripts. **Right Panel.** F_1 scores. Average F_1 scores resulting from 1,000 repeated experiments with vertical bars representing one standard deviation.

Although existing methods can operate in the single-subject, single-sample setting, Table 1 and Fig. 3 demonstrate the quality of inference of iDEG as compared to conventional methods. We note that the comparison is not fair, since these techniques were not explicitly designed for single-subject, single-sample studies – we simply conducted the simulations against these methods to demonstrate the need of new approaches for studying single-subject transcripts in TCWR conditions. In addition, iDEG relies on widely-accepted parametric assumptions of RNA-Seq data; however, more sensitivity studies should be conducted to evaluate the performance when these assumptions are violated. In addition, a true gold standard to evaluate iDEG and other methods is not as simple as obtaining replicates and running conventional methods as pointed out by several recent papers. Conventional methods were recently shown to concordantly agree on biases and false positive results in recent evaluations, suggesting that generating a reliable gold standard for single-subject studies requires extensive strategies to abate and mitigate these issues[32,33]. We intend to leverage the lessons learned in these new studies and build a substantive gold standard from isogenic samples with replicates in numerous conditions.

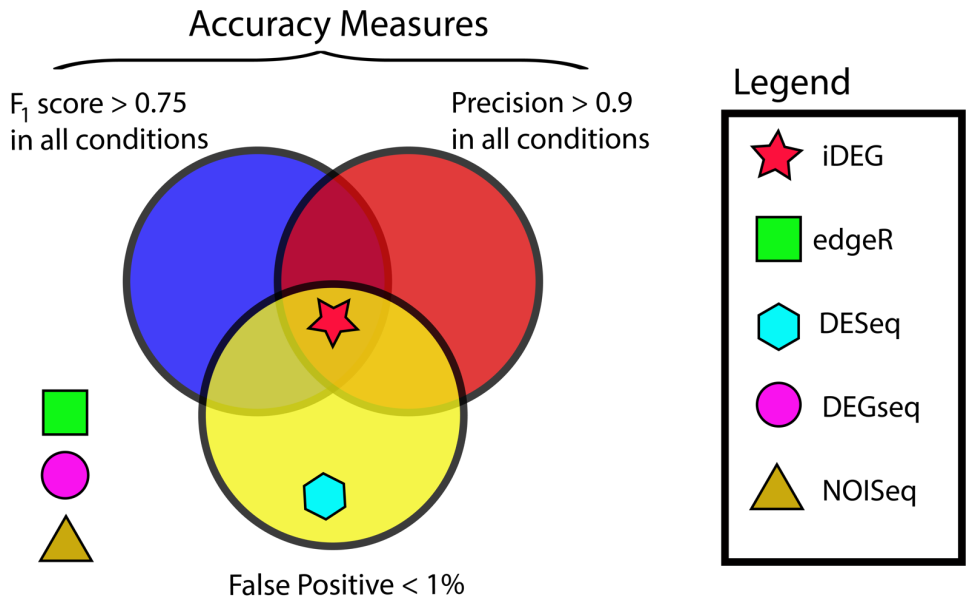

Fig. 4. Although various techniques are resistant to fold change inflation, iDEG is the only technique that also maintains both high precision and recall, demonstrating its value in single-subject studies in TCWR. Note, edgeR and NOISeq attain high F_1 scores when the signal is sufficiently high (~ 4,000 genes seeded as DEG), but struggle to maintain high precision in low-signal settings resulting in a high number of false positives. Methods like DESeq and DEGSeq either detect a very large or small number of DEGs thus resulting in a large number of false positives or false negatives.

5. Conclusion

Over the past decade, state-of-the-art techniques in RNA-Seq data analysis have delivered powerful new tools[14] for extending large-scale inference to small-sample settings. The primary goal of iDEG is not to replace these, but rather to expand the scope of RNA-Seq studies into the single-subject, single-replicate realm. There are myriad techniques available for RNA-Seq analysis[14]. However, very few were designed to produce robust and reliable inference in single-subject, single-sample settings. Given the prohibitive cost and access to clinical tissue, iDEG transcends the replication requirements of conventional statistical methods and fills this niche. By focusing on one patient at a time, iDEG has the potential to ascertain meaningful biomolecular mechanisms for decision-making in precision medicine[34]. iDEG is uniquely poised to do so since it i) calculates a data-driven mean-dispersion relationship, ii) is robust to fold-change inflation, and iii) provides reliable patient-level inference in absence of biological replicates. In future studies, we envision to validate iDEG in a biological framework, and then extend iDEG to identify DEGs under multiple conditions, to identify differentially expressed pathways (by incorporating ontologies and knowledge graphs), to

apply it to other ’omics measures, and to further precision-medicine discoveries in single-subject, single-replicate studies.

References

1. J. Kaiser, *Science*, 2015, DOI: 10.1126/science.aaa6436.
2. N. J. Schork, *Nature*, 2015, **520**, 609-611.
3. E. J. Topol, *Cell*, 2014, **157**, 241-253.
4. M. D. Robinson, D. J. McCarthy and G. K. Smyth, *Bioinformatics*, 2009, **26**, 139-140.
5. S. Anders and W. Huber, *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL)*, 2012.
6. S. Tarazona, F. García, et al., *EMBnet. journal*, 2011, **17**, 18-19.
7. J. Feng, C. A. Meyer, et al., *Bioinformatics*, 2012, **28**, 2782-2788.
8. L. Wang, Z. Feng, et al., *Bioinformatics*, 2009, **26**, 136-138.
9. A. G. Schissler, V. Gardeux, et al., *Bioinformatics*, 2015, **31**, i293-i302.
10. Q. Li, A. G. Schissler, et al., *BMC Medical Genomics*, 2017, **10**, 27.
11. Q. Li, A. G. Schissler, et al., *Journal of biomedical informatics*, 2017, **66**, 32-41.
12. V. Gardeux, I. Achour, et al., *Journal of the American Medical Informatics Association*, 2014, **21**, 1015-1025.
13. S. R. Zaim, Q. Li, et al., 2018.
14. F. Vitali, Q. Li, et al., *Briefings in bioinformatics*, 2017.
15. R. C. Team, 2013.
16. M. D. Robinson and G. K. Smyth, *Biostatistics*, 2007, **9**, 321-332.
17. C. W. Law, Y. Chen, et al., *Genome Biology*, 2014, **15**, R29.
18. S. Anders and W. Huber, *Genome biology*, 2010, **11**, R106.
19. Y. Di, D. W. Schafer, et al., *Statistical Applications in Genetics and Molecular Biology*, 2011, **10**.
20. P. Craven and G. Wahba, *Numerische Mathematik*, 1978, **31**, 377-403.
21. N. F. Laubscher, *Journal of the American Statistical Association*, 1961, **56**, 143-150.
22. F. R. Hampel, *Journal of the American Statistical Association*, 1974, **69**, 383-393.
23. D. C. Montgomery, *Design and analysis of experiments*, John Wiley & Sons, 2008.
24. B. Efron, *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Cambridge University Press, 2012.
25. N. Dean and A. E. Raftery, *BMC bioinformatics*, 2005, **6**, 173.
26. B. Efron and R. Tibshirani, *The Annals of Statistics*, 1996, **24**, 2431-2461.
27. B. Efron, *Journal of the American Statistical Association*, 2004, **99**, 96-104.
28. B. Efron, *The Annals of Statistics*, 2007, **35**, 1351-1377.
29. B. Efron, *Journal of the American Statistical Association*, 2007, **102**, 93-103.
30. B. Efron, R. Tibshirani, et al., *Journal of the American Statistical Association*, 2001, **96**, 1151-1160.
31. Y. Benjamini and Y. Hochberg, *Journal of the royal statistical society. Series B (Methodological)*, 1995, 289-300.
32. J. Costa-Silva, D. Domingues and F. M. Lopes, *PloS one*, 2017, **12**, e0190152.

33. C. Soneson and M. D. Robinson, *Nature methods*, 2018, **15**, 255.
34. V. Gardeux, J. Berghout, et al., *Journal of the American Medical Informatics Association*, 2017, **24**, 1116-1126.