THE UNIVERSITY OF ARIZONA HEALTH SCIENCES
Center for Biomedical Informatics & Biostatistics

Novel statistical methods of single-subject analysis to advance precision medicine

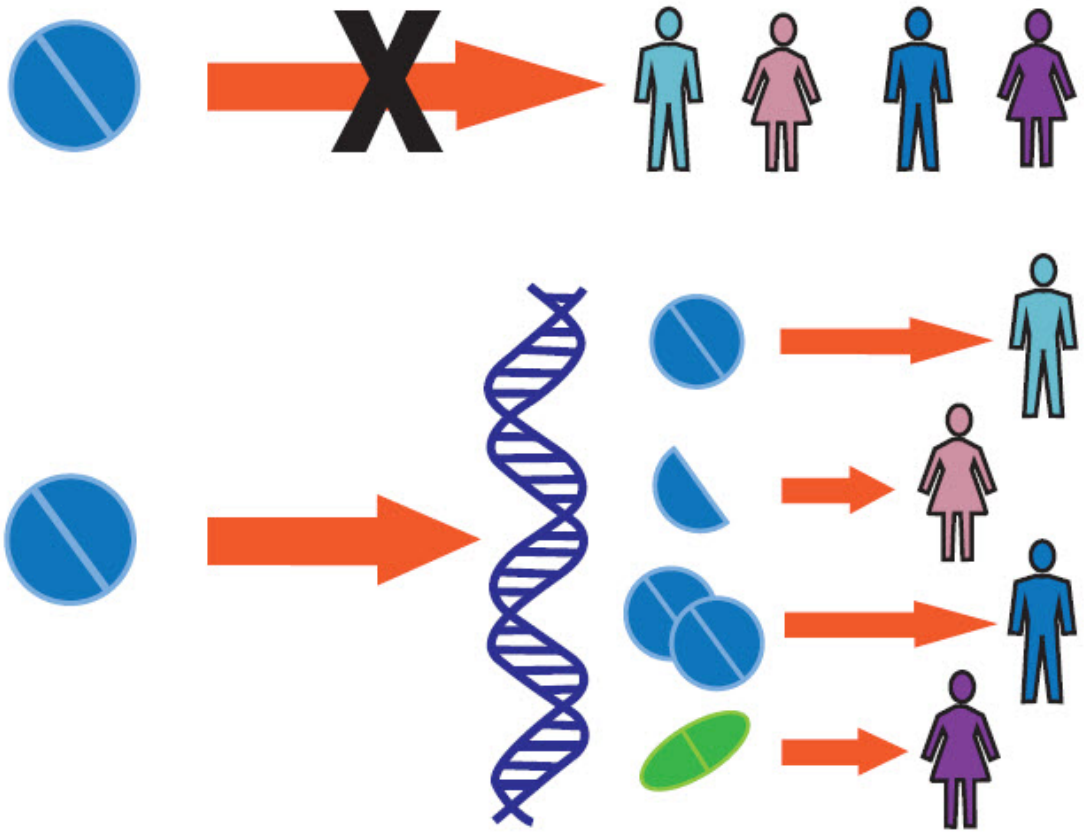
Qike Li

Committee members:
Jin Zhou, PhD
Joseph W. Watkins, PhD
Yves A. Lussier, MD
Hao Helen Zhang, PhD

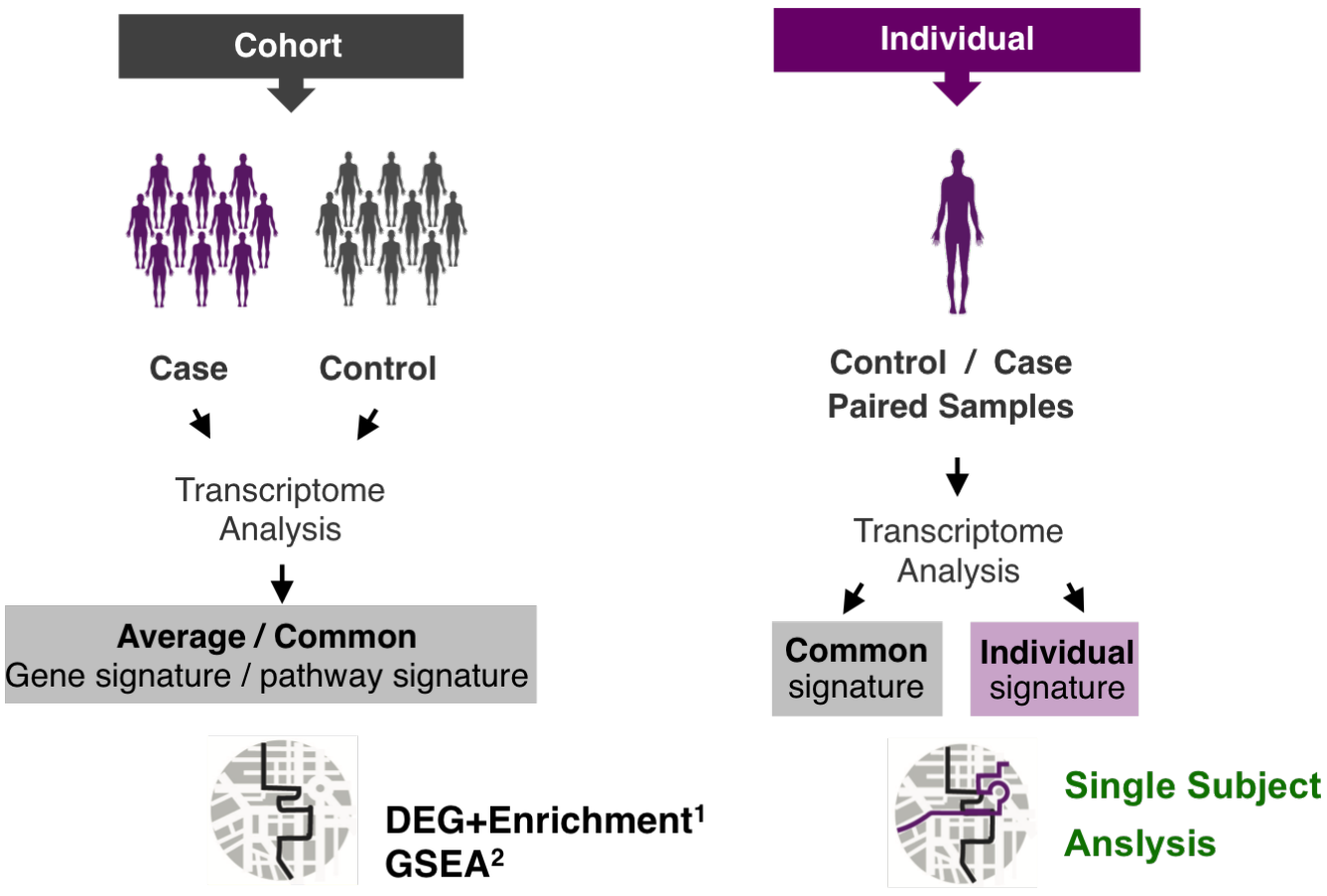
Outline

Outline

Precision Medicine



Precision Medicine



Single-subject analysis: study one patient at a time

- DNA-level
  - Pros: Single-subject analysis has provided groundbreaking insights into the relationship between genetics (DNA-level) variations and diseases[1, 2].
  - Cons: DNA-level variations are only one piece of the puzzle. Environment and the interplays between environment and DNA-level variations also play critical roles in disease development

[1] E. Yong 2013 *National Geographic*.

[2] www.theglobeandmail.com

Single-subject analysis: study one patient at a time

- RNA-level
  - Promise: The impact of DNA-level variations, environment, and the interactions between the two are better captured by the transcriptome. And the differentially expressed genes/mRNAs (DEG) may shed light on disease mechanisms.
  - Challenge: Unlike the static DNA-level variations, the dynamic mRNA expression level is a random variable. Without estimating the variation of this random variable, it is challenging to identify DEG.

Main challenge (Sample size is one)

Example RNA-Seq quantified mRNA expression data

Gene	Case expression	Baseline expression
A1BG	91.96	71.98
A1CF	1.34	0.00
A2BP1	0.33	1.59
A2LD1	127.36	71.38
A2ML1	772.75	11.77
A2M	11824.67	29384.64
A4GALT	892.99	870.93
A4GNT	0.67	5.41
AAA1	0.00	0.00
...	...	...
tAKR	0.00	0.00

Our goal

- How differentially expressed are the genes/gene-sets?
  - An effect size of the genes/gene-sets

Our goal

- How differentially expressed are the genes/gene-sets?
  - An effect size of the genes/gene-sets
- What is the uncertainty of the measurement?
  - A p-value

Our goal

- How differentially expressed are the genes/gene-sets?
  - An effect size of the genes/gene-sets
- What is the uncertainty of the measurement?
  - A p-value
  - A posterior probability

Our solutions

N-of-1 *pathways* MixEnrich

Qike Li, . . . , Hao Helen Zhang, Yves A. Lussier. "N-of-1-pathways MixEnrich: advancing precision medicine via single-subject analysis in discovering dynamic changes of transcriptomes" *BMC Medical Genomics*, Accepted

N-of-1 *pathways* kMEn

Qike Li, . . . , Hao Helen Zhang, Yves A. Lussier. "kMEn: Analyzing noisy and bidirectional transcriptional pathway responses in single subjects." *Journal of biomedical informatics* 66 (2017): 32-41.

iDEG

In progress

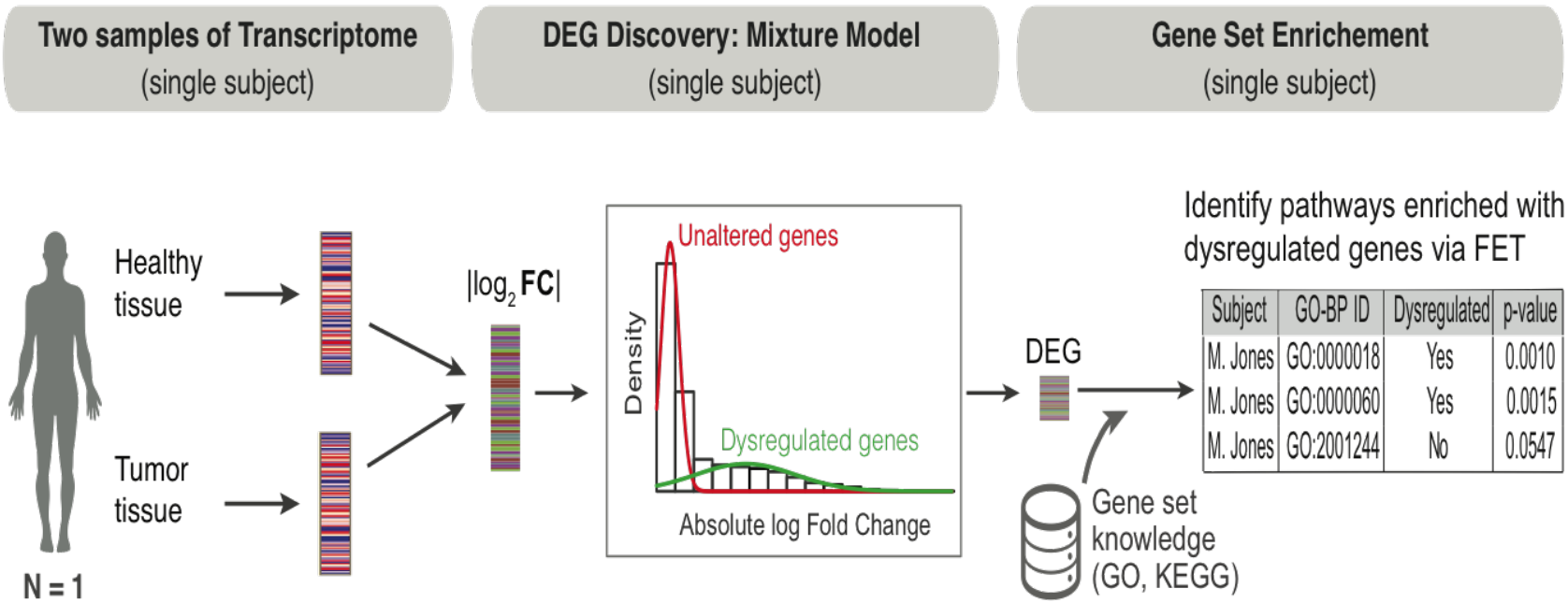
Outline

Our goal

Goal

Identify dysregulated pathways from a pair of transcriptomes(e.g. tumor vs. healthy sample)

Approach


Two samples of Transcriptome
(single subject)
DEG Discovery: Mixture Model
(single subject)
Gene Set Enrichement
(single subject)
Healthy tissue
Tumor tissue
N = 1
|log2 FC|
Unaltered genes
Dysregulated genes
Density
Absolute log Fold Change
DEG
Gene set knowledge (GO, KEGG)
Identify pathways enriched with dysregulated genes via FET
Subject GO-BP ID Dysregulated p-value
M. Jones GO:0000018 Yes 0.0010
M. Jones GO:0000060 Yes 0.0015
M. Jones GO:2001244 No 0.0547

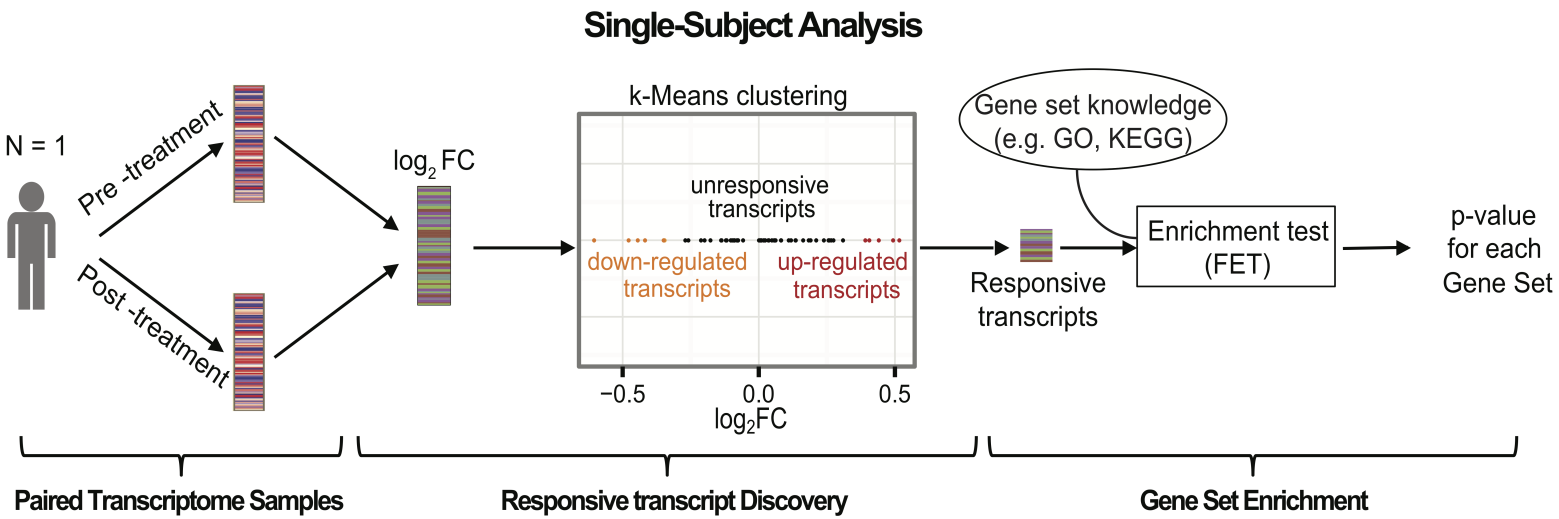
Outline

Our goal

Goal

Provide a robust model under the framework of N-of-1 *pathways*.

Approach


Single-Subject Analysis
N = 1
Pre -treatment
Post -treatment
log2 FC
k-Means clustering
unresponsive transcripts
down-regulated transcripts
up-regulated transcripts
−0.5
0.0
0.5
log2FC
Gene set knowledge (e.g. GO, KEGG)
Responsive transcripts
Enrichment test (FET)
p-value for each Gene Set
Paired Transcriptome Samples
Responsive transcript Discovery
Gene Set Enrichment

Outline

Gaol

Goal

From a pair of transcriptomes collected from a single subject, can you identify differentially expressed genes?

Main Challenges

- For each gene, we have only two numbers two work with.
- Different genes have different variances.

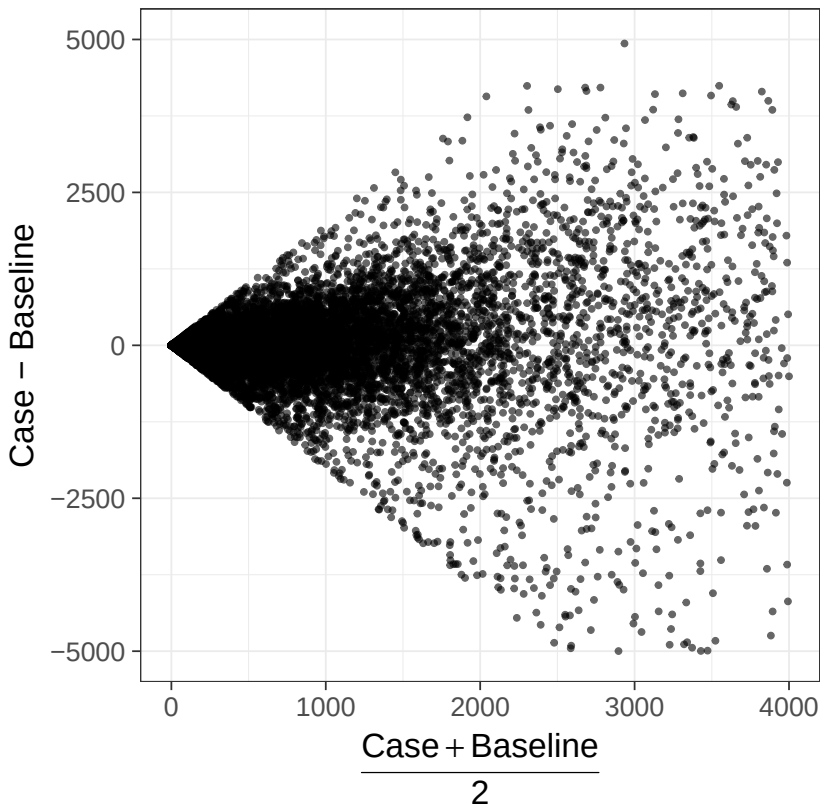
Main Challenges

For each gene, we have only two numbers two work with.

Gene	Case expression	Baseline expression
A1BG	91.96	71.98
A1CF	1.34	0.00
A2BP1	0.33	1.59
A2LD1	127.36	71.38
A2ML1	772.75	11.77
A2M	11824.67	29384.64
A4GALT	892.99	870.93
A4GNT	0.67	5.41
AAA1	0.00	0.00
...	...	...
tAKR	0.00	0.00

Main Challenges

Different genes have different variances.



Existing Methods

- arbitrary cutoff
  - for gene expression fold changes
  - for gene expression absolute differences
- DESeq[1]
  DESeq assumes that most genes are nonDEG and estimates a mean-variance relationship from treating the two samples as if they were replicates.
- edgeR[2]
  edgeR assigns an arbitrary value of the dispersion parameter to all genes and conducts a negative binomial exact test to compute p-values.

Our solution

- Transform RNA-Seq data such that all null genes approximately have the same variance.
- "Borrow strength" across genes and compute the probability of a gene being DEG given the data

$$Pr(\text{gene } g \text{ is a DEG}|Data)$$

Using Poisson distribution to model RNA-Seq data

- While RNA-Seq data are usually modeled by negative binomial distribution to account for over-dispersion of expression counts, the over-dispersion may be negligible and Poisson distribution can fit the data well [1] when the two samples under comparison are processed with due caution.
- Moreover, the assumption of Poisson distribution facilitates a simpler testing procedure of iDEG.

simulate single-subject dataset

We simulate a pair of transcriptomes as an illustrative example

$$
\begin{aligned}
Y_{g1} &\sim Poisson(\mu_{g1}) \\
Y_{g2} &\sim Poisson(\mu_{g2}) \\
P(\mu_{g1}) &= \frac{1}{|\mathcal{B}|} \\
&\mu_{g1} \in \mathcal{B} \quad g = 1, 2, \cdots, G = 20000 \\
&\text{where } \mathcal{B} = \{10, 11, \cdots, 10000\}
\end{aligned}
$$

simulate single-subject dataset

$$
\begin{aligned}
\mu_{g2} &= \begin{cases} \mu_{g1} & \text{if } g \in \overline{\mathcal{G}}, \\ d^s \mu_{g1} & \text{if } g \in \mathcal{G}. \end{cases} \\
\frac{|\mathcal{G}|}{|\mathcal{G}| + |\overline{\mathcal{G}}|} &= 0.1 \\
d &= \frac{\mu_{g1} + n\sqrt{\mu_{g1}}}{\mu_{g1}} \\
s &= \begin{cases} 1 & \text{with probability of } 0.5, \\ -1 & \text{with probability of } 0.5. \end{cases} \\
n &\sim \mathcal{N}(4, 1)
\end{aligned}
$$

where elements in $\mathcal{G}$ are randomly sampled from $\{1, 2, \cdots, 20000\}$ without replacement.

simulate single-subject dataset

Variance Stabilizing Transformation (VST)

Freeman and Tukey[1] developed a variance-stabilizing transformation (VST), $h_{Pois}(\cdot)$, to transform Poisson data, such that the transformed data approximately follow a Normal distribution with the same variance regardless of their means. Namely, if $Y_{gd} \sim Poisson(\mu_{gd})$, then

$$h_{Pois}(Y_{gd}) = \sqrt{Y_{gd}} + \sqrt{Y_{gd}+1} \dot{\sim} N(\mu = \sqrt{\mu_{gd}} + \sqrt{\mu_{gd}+1}, \sigma^2 = 1)$$
$$g = 1, \cdots, G; d = 1, 2$$

Variance Stabilizing Transformation (VST)

$$D_g^* = \sqrt{Y_{g1}} + \sqrt{Y_{g1}+1} - (\sqrt{Y_{g2}} + \sqrt{Y_{g2}+1}) \dot{\sim} N(\mu = 0, \sigma = \sqrt{2})$$

$$g \in \overline{\mathcal{G}}$$

Variance Stabilizing Transformation (VST)

Variance Stabilizing Transformation (VST)

This procedure, taking the difference of VST transformed RNA-Seq data for each gene, makes d_g^* of all null genes approximately follow the same distribution. We now seek to compute the probability of a gene being differentially expressed given the observation of d_g^*:

$$Pr(\text{gene } g \text{ is a DEG} | d_g^*)$$

Local false positive rate

Efron[1, 2, 3] published a series of work on local false discovery rate (*fdr*), the posterior probability of a null hypothesis being true given a summary statistic, in the context of large-scale parallel inference.

$$fdr \equiv Pr(null|\text{summary statistic}) = 1 - Pr(\text{gene } g \text{ is a DEG}|d_g^*)$$

[3] Efron, B. 2007, *The Annals of Statistics*

Local fdr–Two-group model

We assume the summary statistics D_g^* of G genes from a mixture of two distributions, and use empirical Bayesian approach to estimate the posterior probability of genes being null given D_g^*. We first standardize D_g^*,

$$Z_g = \frac{D_g^*}{MAD(D^*)}$$

$$\text{where } MAD(D^*) = (|D_g^* - (D^*)|), \quad g = 1, \cdots, G$$

Local fdr–Two-group model

Suppose G genes are measured in the RNA-Seq experiment, each of the genes is either null or differentially expressed with prior probabilities π_0 or $\pi_1 = 1 - \pi_0$. And the density function of z_g is either $f_0(z)$ or $f_1(z)$.

$$\pi_0 = Pr\{\text{gene } g \text{ is null}\} \quad \text{density is } f_0(z) \text{ if null}$$
$$\pi_1 = Pr\{\text{gene } g \text{ is DEG}\} \quad \text{density is } f_1(z) \text{ if DEG}$$

The marginal mixture density is:

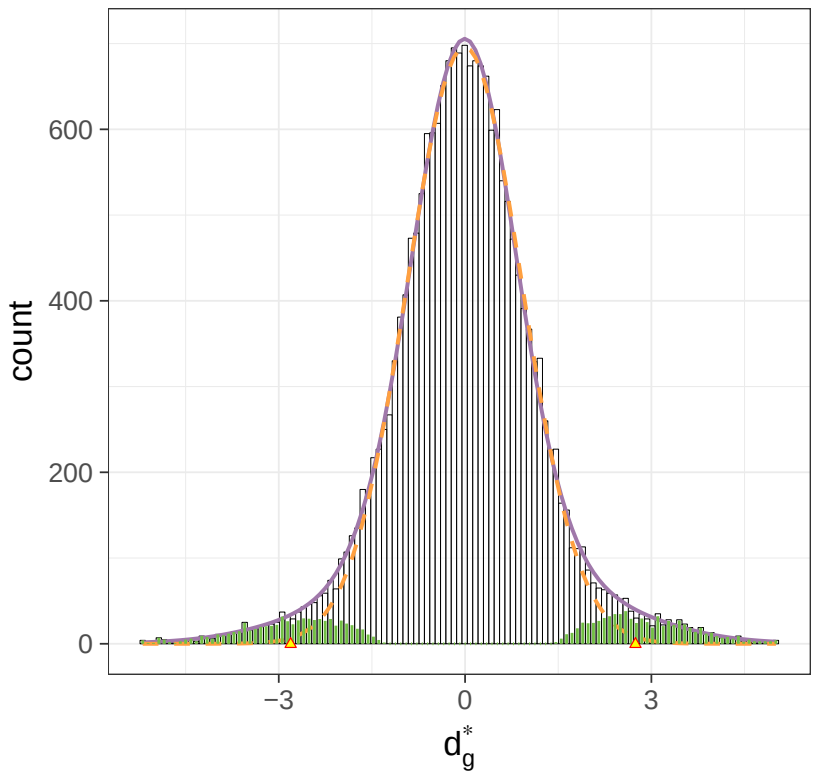
$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$$

Local fdr–Two-group model

Then the local false positive rate, fdr, is the Bayes posterior probability that a gene g is a null gene given z_g:

$$fdr(z) \equiv Pr\{\text{null gene}|z\} = \pi_0 f_0(z)/f(z)$$

Local fdr–Marginal distribution estimation



Local fdr–Marginal distribution estimation

We estimate the mixture density $f(z)$ with smooth but flexible parametric models.
We assume Z belongs to a K-parameter exponential family with natural parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_K)'$ and sufficient statistic $\boldsymbol{x} = (z, z^2, \cdots, z^K)'$

$$f(z) = exp\left\{ \sum_{k=0}^{K} \beta_k z^k \right\}$$

Local fdr–Marginal distribution estimation

Lindsey's method [1] supposes G $z_g's$ have been binned into M bins, giving the counts of $N = (n_1, n_2, \cdots, n_m)'$ of z.

$$\begin{aligned} \mathcal{Z} &= \cup_{m=1}^{M} \mathcal{Z}_m \\ n_m &= \#\{z_g \in \mathcal{Z}_m\} \end{aligned}$$

Local fdr–Marginal distribution estimation

The counts n_m amount to the counts in the bins of the histogram. Now the density $f_\beta(z)$ is reduced to

$$N \sim \text{multinomial}_M(G, \boldsymbol{\pi})$$

where $\pi_m(\beta) = w f_\beta(x_m)$. And w is the bin width in the histogram; x_m is midpoint of z_g in $\mathcal{Z}_m$.

Local fdr–Marginal distribution estimation

Based on the relationship between multinomial and Poisson distribution, we assume the n_m to be the independent Poisson observations.

$$n_m \overset{ind}{\sim} Pois(v_m), \quad v_m = G\pi_m(\beta) \qquad k = 1, 2, \cdots, K$$

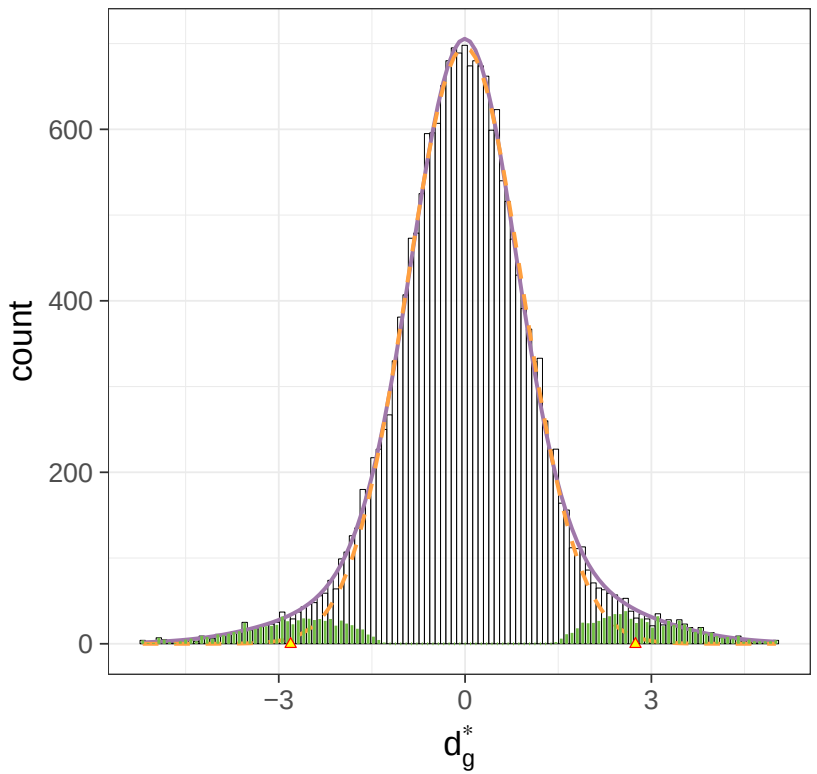
Local fdr–Marginal distribution estimation

Then we use standard Poisson generalized linear model to compute $\hat{\boldsymbol{\beta}}$

$$log(\boldsymbol{v}) = \alpha_0 + \boldsymbol{X}\boldsymbol{\beta}$$

where X is the design matrix with m^{th} row as $(x_m, x_m^2, \cdots, x_m^K)'$.

Local fdr–Empirical null distribution estimation



Local fdr–Empirical null distribution estimation

In large-scale simultaneous hypothesis testing, the theoretical null may be deficient due to various reasons: correlation across genes, correlation between RNA-Seq samples, unobserved covariates (e.g. gender, age, smoking status, etc.), and failed mathematical assumptions (e.g. asymptotic approximation). Fortunately, in large-scale simultaneous testing, the parallel structure allows the estimation of an empirical null distribution, via empirical Bayes, from the own data of the study.

Local fdr–Empirical null distribution estimation

Without assumptions of the parametric forms of of f_0 and f_1, the two-group model is unidentifiable. To restore the identifiability, we make the "zero assumption", which enables the estimation of π_0 and f_0.

$$f_1(z) = 0 \qquad \text{for } z \in \mathcal{A}_0$$

Local fdr–Empirical null distribution estimation

$$\begin{aligned} \mathcal{G}_0 &= \{g : z_g \in \mathcal{A}_0\} \\ N_0 &= |\mathcal{G}_0| \\ \boldsymbol{z}_0 &= \{z_g, g \in \mathcal{G}_0\} \\ f_0(z) &= \frac{1}{\sigma_0}\varphi\left(\frac{z-\mu_0}{\sigma_0}\right) \end{aligned}$$

$$f_{\mu_0,\sigma_0,p_0}(\boldsymbol{z_0}) = \left[\binom{G}{N_0} p^{N_0}(1-p)^{G-N_0}\right]\left[\prod_{\mathcal{G}_0} \frac{\varphi_{\mu_0,\sigma_0}(z_g)}{\int_{\mathcal{A}_0}\varphi_{\mu_0,\sigma_0}(z)dz}\right]$$

where

$$p = \pi_0 \int_{\mathcal{A}_0} \varphi_{\mu_0,\sigma_0}(z)dz = Pr\{z_g \in \mathcal{A}_0\}$$

Local fdr–Empirical null distribution estimation

In consequence,

$$\widehat{\pi_0} = \frac{\hat{p}}{\int_{\mathcal{A}_0} \hat{f}_0(z)dz}$$

Local false positive rate (local fdr)

Finally, with the estimate of marginal distribution, $\hat{f}$ from section , the estimate of $\hat{\pi}_0$, and the estimate of empirical null density, $\hat{f}_0$, we get to compute $\widehat{fdr}$

$$\widehat{fdr}(z) = \widehat{\pi_0}\hat{f}_0(z)/\hat{f}_(z)$$

Procedure

- Step 1: Simulate one single-subject datasets:

$$
\begin{aligned}
Y_{g1} &\sim Poisson(\mu_{g1}) \\
Y_{g2} &\sim Poisson(\mu_{g2}) \\
P(\mu_{g1}) &= \frac{1}{500} e^{-\frac{1}{500} \times \mu_{g1}} \qquad g = 1, \cdots, 20\,000
\end{aligned}
$$

Procedure

$$
\begin{aligned}
\mu_{g2} &= \begin{cases} \mu_{g1} & \text{if } g \in \overline{\mathcal{G}}, \\ d^s \mu_{g1} & \text{if } g \in \mathcal{G}. \end{cases} \\
\frac{|\mathcal{G}|}{|\mathcal{G}| + |\overline{\mathcal{G}}|} &= p; \quad p = 0.05 \\
d &= \frac{\mu_{g1} + n\sqrt{\mu_{g1}}}{\mu_{g1}} \\
s &= \begin{cases} 1 & \text{with probability of } 0.5, \\ -1 & \text{with probability of } 0.5. \end{cases} \\
n &\sim \mathcal{N}(9, 1)
\end{aligned}
$$

where elements in $\mathcal{G}$ are randomly sampled from $\{1, 2, \cdots, 20000\}$ without replacement.

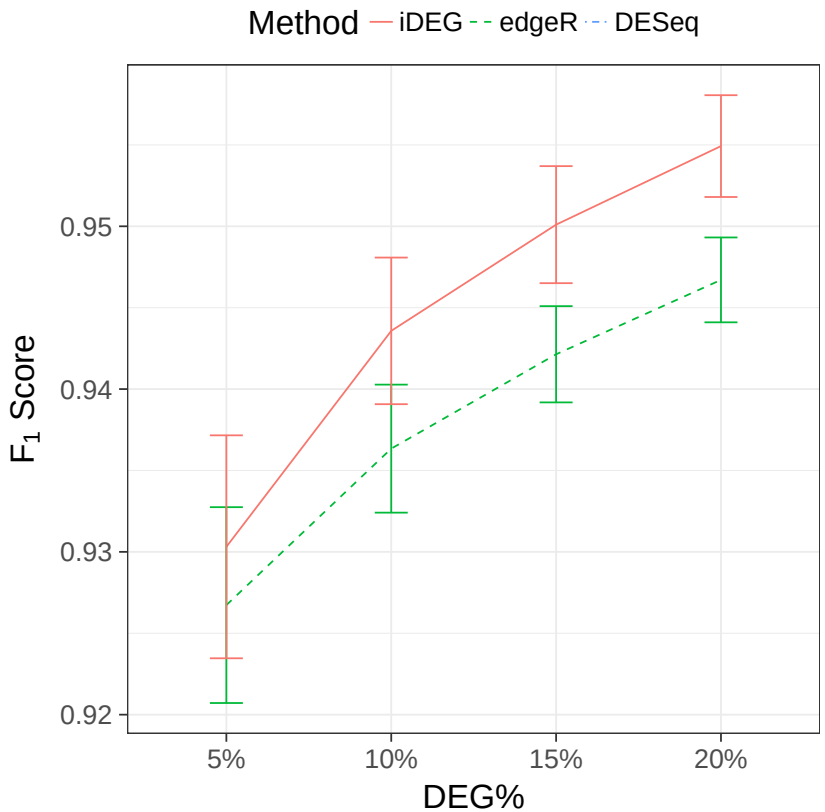
Procure

- Step 2: Conduct iDEG, DESeq, and edgeR.
- Step 3: Compute F1 score for each method,

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

- Step 4: Repeat Step1-Step3 for 1000 times
- Step 5: Calculate the arithmetic mean and standard deviation of the 1000 F_1 resulted from each method.
- Step 6 Change the value of p, repeat Step 1-Step 5

Method evaluation (Poisson case)



Method evaluation (Poisson case)

DEG%	Method	Precision	Recall (TPR)	FPR	F1	Predictions
5%	iDEG	0.987 (4.8e-03)	0.879 (1.3e-02)	0.001 (2.3e-04)	0.929 (7.1e-03)	890.46 (1.5e+01)
	edgeR	0.918 (8.2e-03)	0.934 (8.3e-03)	0.004 (4.8e-04)	0.926 (6.2e-03)	1017.52 (1.2e+01)
	DESeq	NaN (NA)	0 (0.0e+00)	0 (0.0e+00)	NaN (NA)	0 (0.0e+00)
10%	iDEG	0.988 (3.3e-03)	0.904 (8.1e-03)	0.001 (3.4e-04)	0.944 (3.9e-03)	1829.1 (2.0e+01)
	edgeR	0.923 (5.2e-03)	0.95 (5.1e-03)	0.009 (6.5e-04)	0.936 (3.4e-03)	2059.95 (1.7e+01)
	DESeq	NaN (NA)	0 (0.0e+00)	0 (0.0e+00)	NaN (NA)	0 (0.0e+00)
15%	iDEG	0.991 (2.4e-03)	0.913 (7.5e-03)	0.001 (4.0e-04)	0.95 (3.6e-03)	2764.26 (2.7e+01)
	edgeR	0.926 (5.0e-03)	0.959 (3.4e-03)	0.014 (1.0e-03)	0.942 (3.0e-03)	3105.66 (2.1e+01)
	DESeq	NaN (NA)	0 (0.0e+00)	0 (0.0e+00)	NaN (NA)	0 (0.0e+00)
20%	iDEG	0.991 (2.3e-03)	0.921 (6.3e-03)	0.002 (5.4e-04)	0.955 (2.9e-03)	3716.23 (3.1e+01)
	edgeR	0.93 (4.1e-03)	0.963 (3.1e-03)	0.018 (1.1e-03)	0.946 (2.7e-03)	4143.99 (2.2e+01)
	DESeq	NaN (NA)	0 (0.0e+00)	0 (0.0e+00)	NaN (NA)	0 (0.0e+00)

Although the Recall/TPR and number of precisions of iDEG are lower than edgeR, iDEG has high precision and low FPR across all percentages of DEG. These operating characteristics of iDEG may be preferable in large-scale inference, like RNA-Seq analysis, where investigators examines tens of thousands of genes in a high-throughput manner.

Negative binomial distribution

Parameter δ_g in negative binomial distribution accounts for the overdispersion.

$$
\begin{aligned}
P(y_{gd}|\mu_{gd},\delta_g) &= (1+\delta_g\mu_{gd})^{-1/\delta_g}\frac{\Gamma(y_{gd}+1/\delta_g)}{y_{gd}!\Gamma(1/\delta_g)}\left(\frac{\delta_g\mu_{gd}}{1+\delta_g\mu_{gd}}\right)^{y_{gd}} \\
& \qquad y_{gd}=0,1,\cdots \quad g=1,\cdots,G; \quad d=1,2 \\
E(Y_{gd}) &= \mu_{gd} \\
Var(Y_{gd}) &= \mu_{gd}+\delta_g\mu_{gd}^2
\end{aligned}
$$

VST for negative binomial distribution

$$h_{nb}(Y_{gd}) = \frac{1}{\sqrt{\delta_g}} sinh^{-1}\sqrt{Y_{gd}\delta_g} \dot{\sim} N(\mu = \frac{1}{\sqrt{\delta_g}} sinh^{-1}\sqrt{\mu_{gd}\delta_g}, \sigma^2 = \frac{1}{4})$$

Assumptions and approximation

- We assume $\delta_{g1} = \delta_{g2} = \delta_g$ and δ_g is a smooth function of the expression mean μ_g. Consequently, genes with the same μ_g follow the same distribution.
- We make an approximation by pooling genes with close expression means to estimate δ_g. Specifically,

$$Y_{gd} \dot{\sim} NB(\mu_w, \delta_w)$$
$$\forall g \in \{g : Gene_g \in \{\text{genes in } w^{th}\text{window}\} \cap \{nonDEG\}\}$$
$$d = 1, 2$$

- We assume the majority of the genes are null genes, and estimate μ_w and σ_w^2 for each window by outlier robust estimators, median and MAD, respectively.

smooth spline

After estimating $\hat{\mu}_w$ and $\hat{\delta}_w$ from each window, we fit a smooth spline of $\hat{\mu}_w$ and $\hat{\delta}_w$ by minimizing the penalized residual sum of squares:

$$RSS(f, \mu) = \sum_{w=1}^{W} (\hat{\delta}_w - f(\hat{\mu}_w))^2 + \mu \int f''(\hat{\mu}_w)^2 dt$$

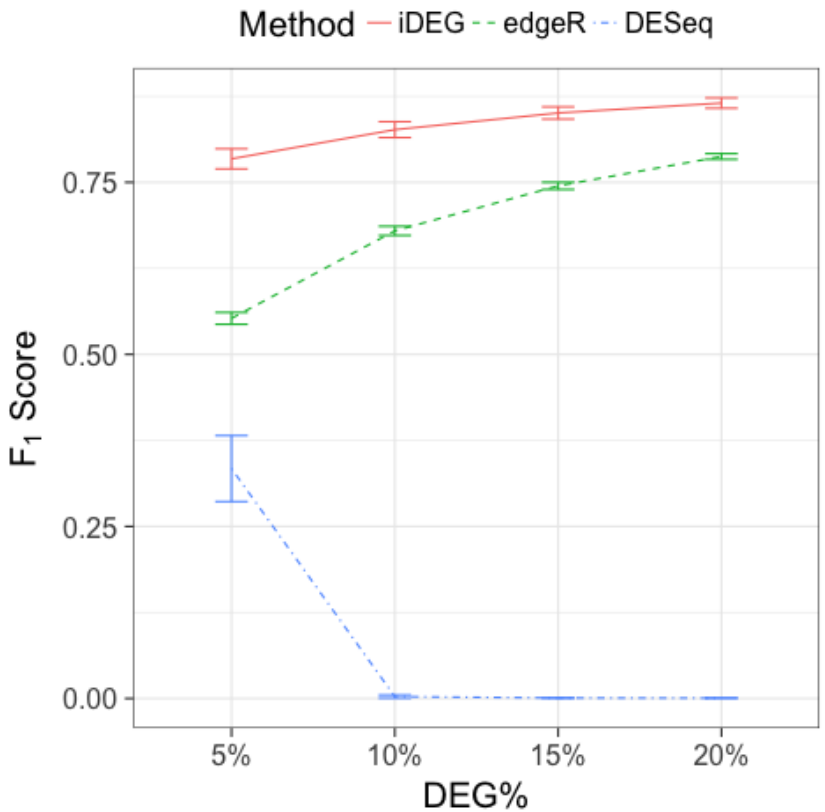
Numerical study (negative binomial case)

Data simulation

Same as the procedure in the Poisson case except,

$$\begin{aligned} Y_{g1} &\sim NB(\mu_{g1}, \delta_g) \\ Y_{g2} &\sim NB(\mu_{g2}, \delta_g) \\ \delta_g &= 0.005 + 9/(\mu_{g1} + 100) \end{aligned}$$

Numerical study (negative binomial case)



Numerical study (negative binomial case)

DEG..	Method	Precision	Recall.TPR	FPR	F1	Predictions
5%	iDEG	0.93 (1.6e-02)	0.679 (2.5e-02)	0.003 (7.2e-04)	0.784 (1.5e-02)	730.482 (3.4e+01)
	edgeR	0.39 (8.4e-03)	0.948 (7.1e-03)	0.078 (2.8e-03)	0.552 (8.6e-03)	2432.95 (5.4e+01)
	DESeq	1 (1.3e-03)	0.202 (3.5e-02)	0 (1.4e-05)	0.334 (4.8e-02)	201.589 (3.5e+01)
10%	iDEG	0.946 (9.8e-03)	0.734 (2.1e-02)	0.005 (9.8e-04)	0.827 (1.2e-02)	1552.089 (5.5e+01)
	edgeR	0.527 (7.7e-03)	0.956 (4.7e-03)	0.095 (2.9e-03)	0.68 (6.6e-03)	3628.589 (5.4e+01)
	DESeq	1 (0.0e+00)	0.001 (1.0e-03)	0 (0.0e+00)	0.003 (2.3e-03)	1.107 (2.0e+00)
15%	iDEG	0.955 (7.1e-03)	0.767 (1.6e-02)	0.006 (1.1e-03)	0.851 (8.9e-03)	2409.492 (6.4e+01)
	edgeR	0.608 (6.8e-03)	0.96 (3.6e-03)	0.109 (3.1e-03)	0.745 (5.2e-03)	4735.157 (5.5e+01)
	DESeq	1 (0.0e+00)	0 (4.8e-05)	0 (0.0e+00)	0.001 (1.1e-04)	0.02 (1.4e-01)
20%	iDEG	0.961 (5.8e-03)	0.787 (1.4e-02)	0.008 (1.3e-03)	0.865 (7.5e-03)	3275.416 (7.3e+01)
	edgeR	0.666 (5.7e-03)	0.964 (3.0e-03)	0.121 (3.1e-03)	0.788 (4.1e-03)	5791.611 (5.2e+01)
	DESeq	1 (0.0e+00)	0 (7.9e-06)	0 (0.0e+00)	0 (0.0e+00)	0.001 (3.2e-02)

Outline

Future work

- Build an R package for iDEG.
- Submit the manuscript.

Acknowledgements

Committee members

- Jin Zhou, PhD
- Joseph W. Watkins, PhD
- Yves A. Lussier, MD
- Hao Helen Zhang, PhD

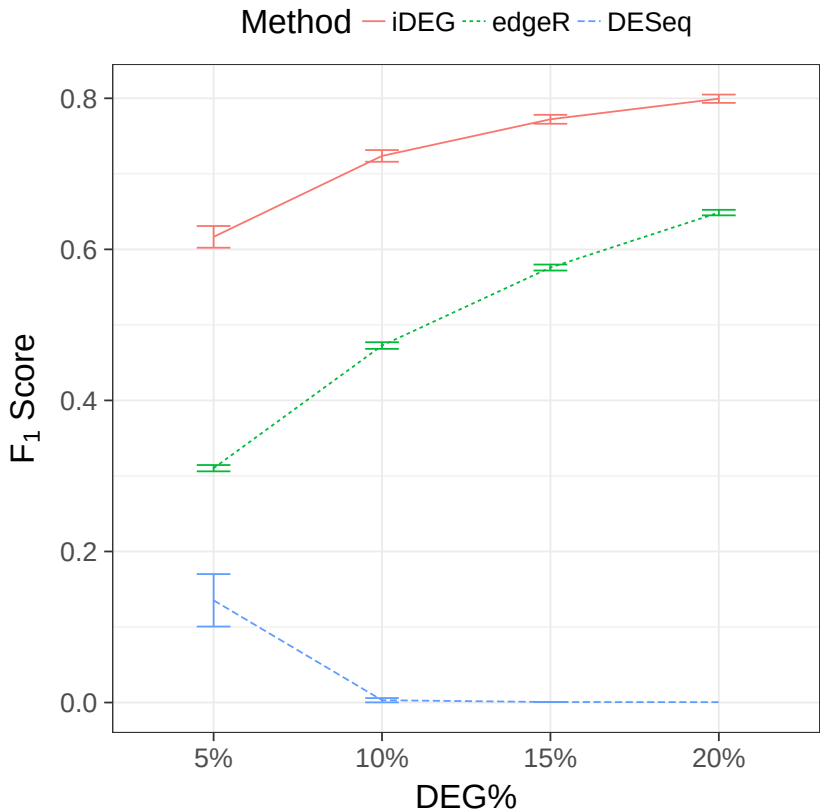
Sensitivity Analysis

Our experience indicates that without making assumptions on RNA-Seq data, it is difficult to construct suitable statistical inferences for a single-subject dataset. Two main assumptions we make are: 1) the value of dispersion parameter is a function of expression mean, and 2) the majority of the genes are null genes. In spite of the prevalence of these assumptions in literature, we examined the sensitives of iDEG to these assumptions.

iDEG is robust to the assumption that dispersion is a function of expression mean

We simulated negative binomial distributed RNA-Seq data with the values of dispersion parameter δ_g from a uniform distribution $unif(0.001, 0.1)$. Despite the performance drop of both iDEG and edgeR, iDEG performed reasonably well, and its F_1 scores were the highest among the three methods.

iDEG is robust to the assumption that dispersion is a function of expression mean



iDEG is robust to the assumption that the majority genes are null genes

The sensitivity of iDEG to this assumption was tested by comparing the three methods on simulated single-subject datasets with a series of percentages of DEG. Although edgeR doesn't make the this assumption, until the percentage of DEG reaches 40%, iDEG still performs better than edgeR. Note, 40% DEG is an unrealistic extreme case in biology.

iDEG is robust to the assumption that the majority genes are null genes