Statistical Methods for RNA-Seq Data

Tina Shi

Advisor: Dr. Joshua Habiger

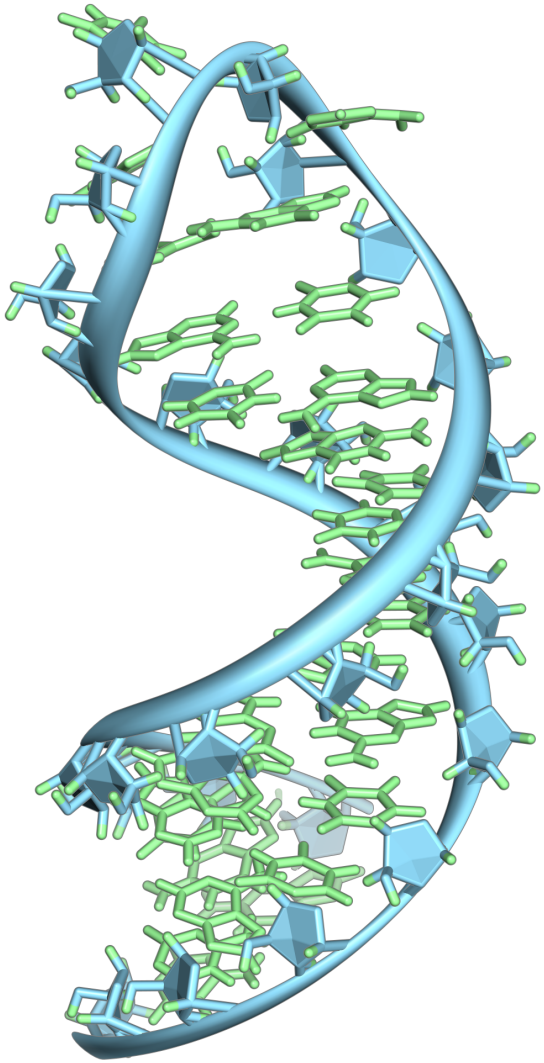
Oklahoma State University

July 3, 2019

Outline

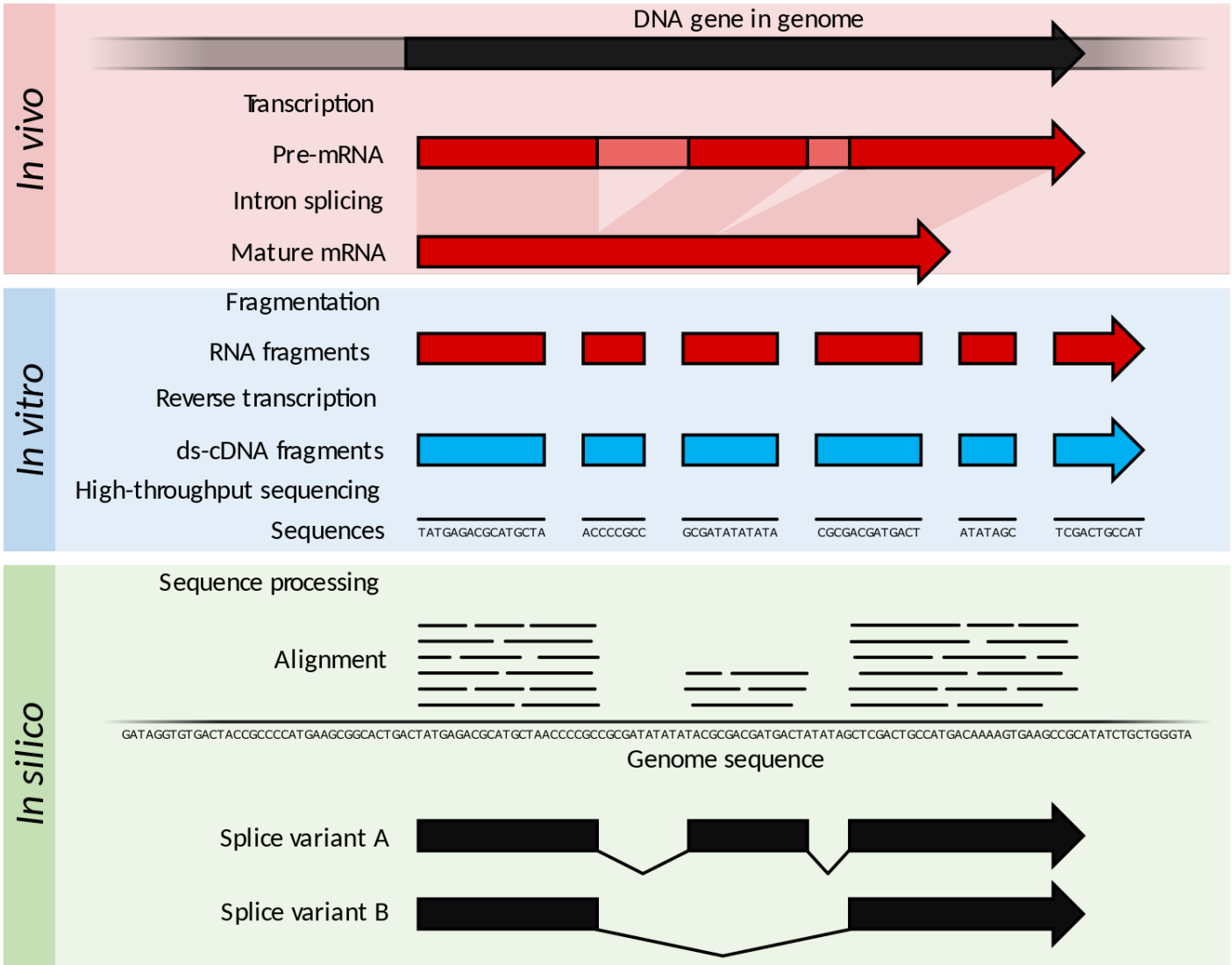
RNA

Nucleobases:
guanine(G),
uracil(U),
adenine(A), and
cytosine(C).



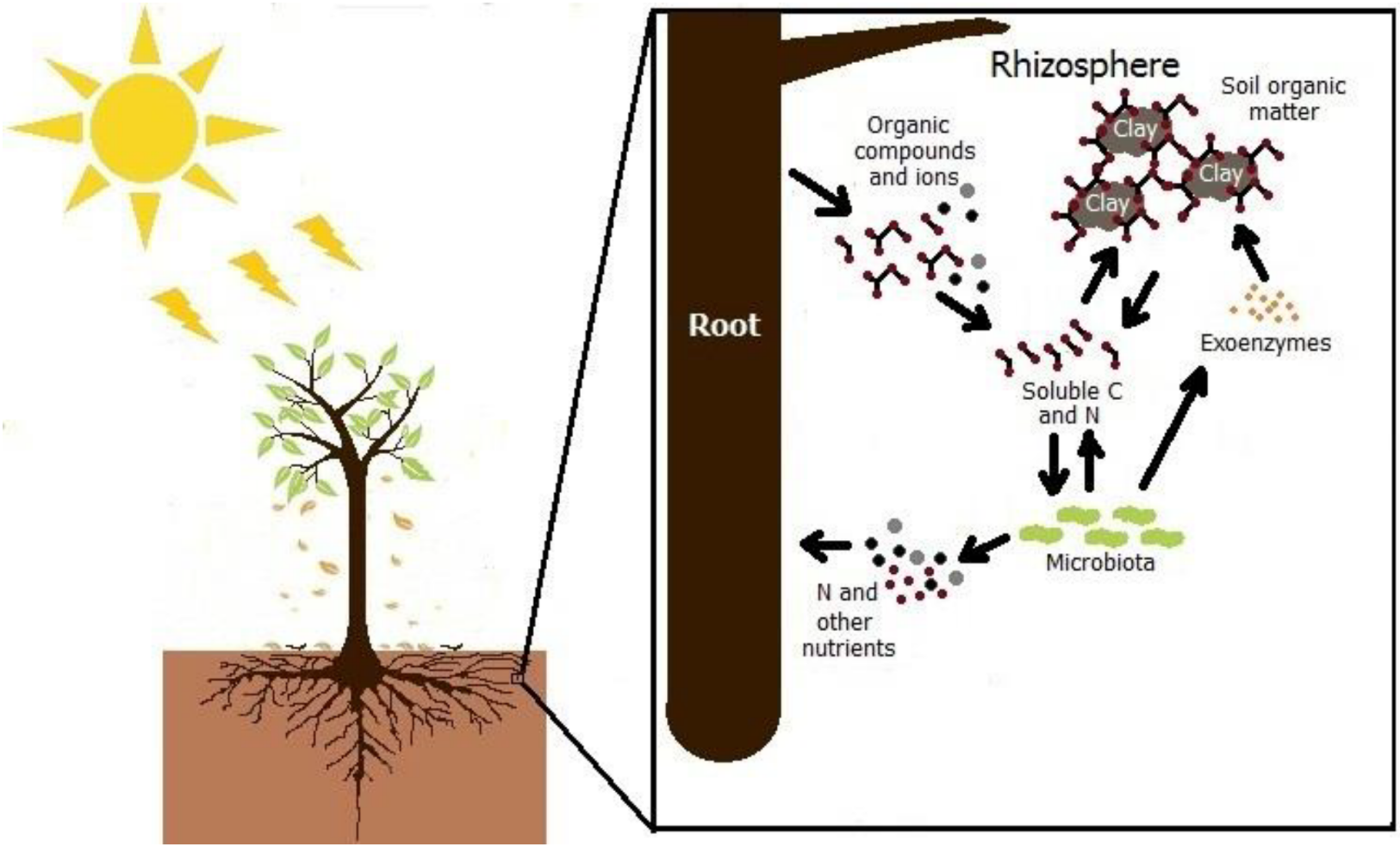
[2] "Pre-mRNA-1ysv-tubes" is licensed by Vossman under CC BY-SA 3.0.

RNA-sequencing (RNA-Seq)



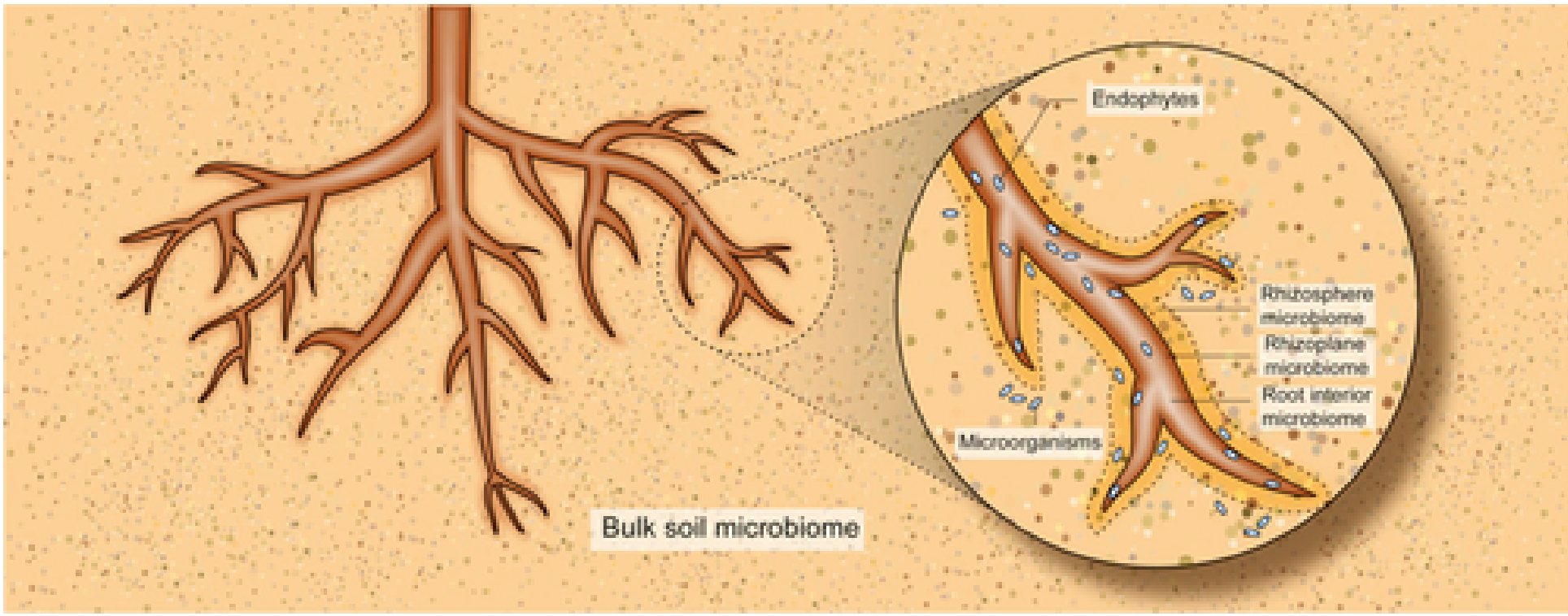
[2] "Summary of RNA-Seq" is licensed by Thomas Shafee under CC BY 4.0.

Rhizosphere

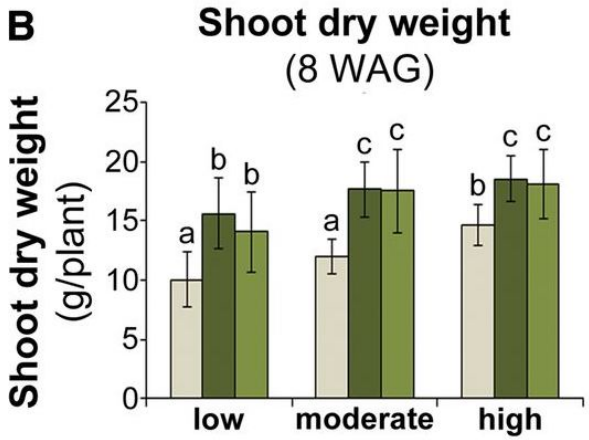


Microbiome

Microbiome includes bacteria, fungi, viruses, ...



Shoot Biomass



RNA-Seq data

Shoot biomass $\boldsymbol{x} = (x_1, \ldots, x_9) = (7.58, \ldots, 10.53)$

Operational taxonomic units (OTUs): grouped similar 16S rRNA sequences

OTU m	y_{m1}	y_{m2}	$\cdots$	y_{m9}
1	4230	3563	$\cdots$	1954
2	3523	3222	$\cdots$	1559
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
1592	0	2	$\cdots$	0

Hypothesis Test

Assume $Y_{mn} \sim \text{Poisson}(\mu_{mn})$ and $\log(\mu_{mn}) = \alpha_m + \beta_m x_n$.

$H_m : \mu_{m1} = \cdots = \mu_{m9}$ or $H_m : \beta_m = 0$.

$$z_m = \frac{\hat{\beta}_m}{I_\beta^{-1/2}} \tag{1}$$

$$p_m = 2[1 - \Phi(|z_m|)] \tag{2}$$

p-value Method

Type I error (false positive or false discovery): H_m is rejected when it is true.

Reject H_m when $p_m \leq \alpha$.

$\alpha M \to \infty$ as $M \to \infty$

- Suppose $\alpha = 0.05$, what if $M = 100$?
- What if $M = 1500$?

False Discovery Rate

	Non-rejected null	Rejected null	Total
True null	U	V	M_0
True non-null	T	S	$M - M_0$
Total	$M - R$	R	M

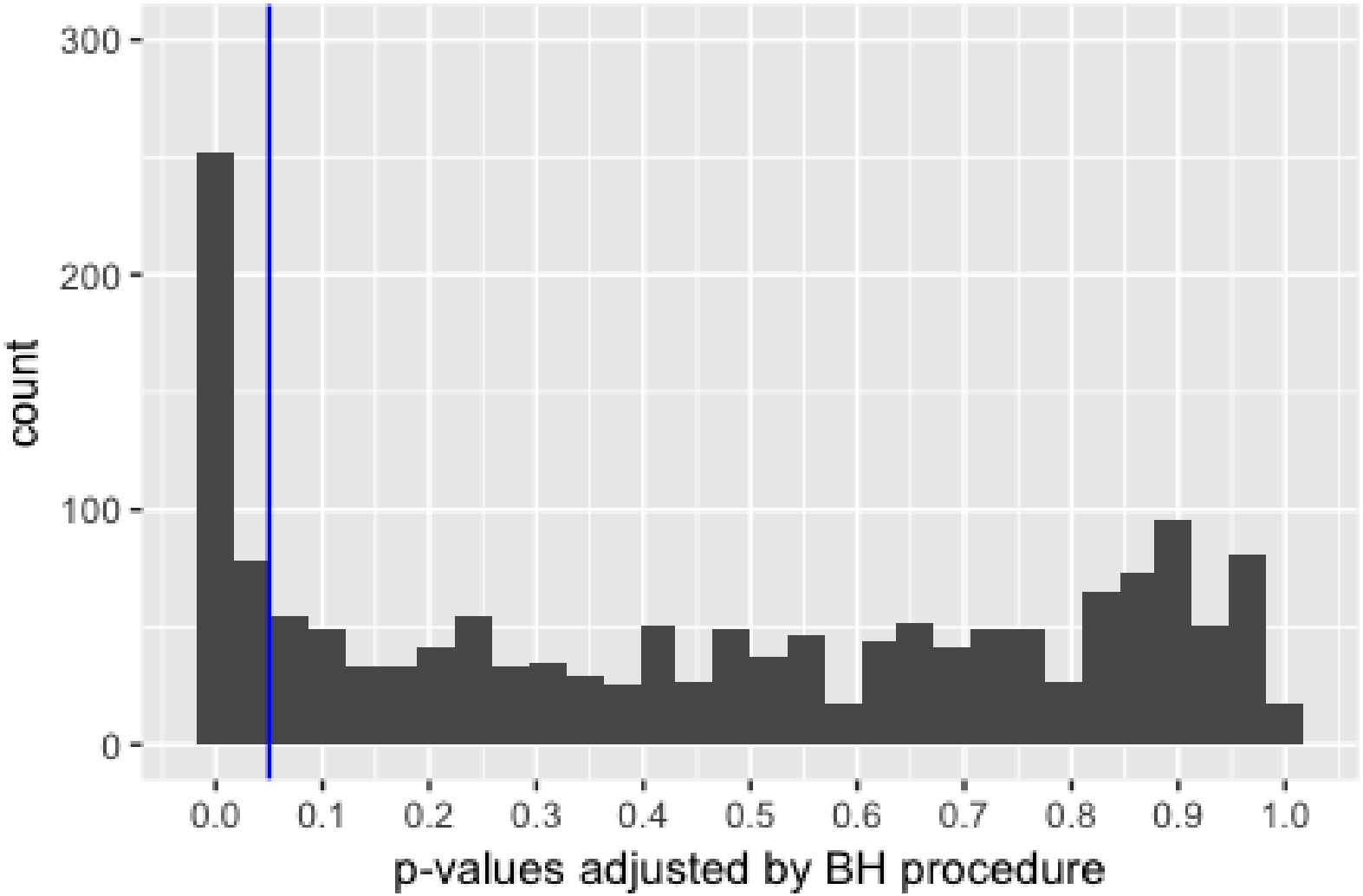
$$FDR = E\left(\frac{V}{R}|R \neq 0\right)$$

BH procedure (Benjamini and Hochberg, 1995):
Reject j null hypotheses with smallest adjusted p-values:

$$j = \max\{m : \frac{M}{m} P_{(m)} \leq \alpha\}.$$

BH Procedure Result

328 ($\sim$ 20.6%) OTUs are discovered at $\alpha = 0.05$.



Local False Discovery Rate

$p_0 = P(null)$ $f_0(z)$ density if null

$p_1 = P(nonnull)$ $f_1(z)$ density if nonnull

Define the mixture density

$$f(z) = p_0 f_0(z) + p_1 f_1(z) \tag{3}$$

Define the local false discovery rate (IFDR) ($IFDR$)(Efron, 2008)

$$IFDR(z) = P(null|Z = z) = \frac{p_0 f_0(z)}{f(z)}$$

IFDR procedure

1. Calculate *IFDR*

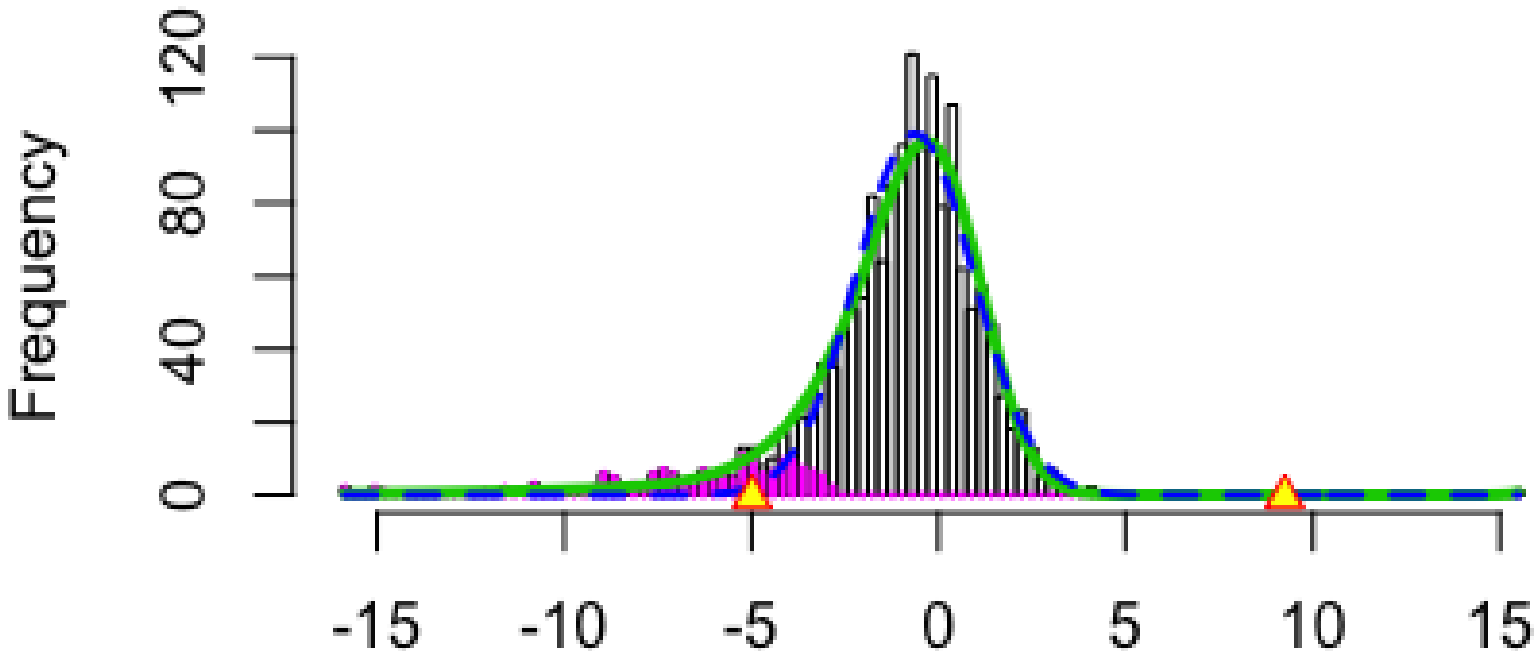
$$IFDR(z) = P(null|Z = z) = \frac{p_0 f_0(z)}{f(z)}$$

2. Reject j null hypotheses with smallest *IFDR*:

$$j = \max\{m : \frac{\sum_{i=1}^{m} IFDR_{(i)}}{m} \leq \alpha\}$$

IFDR Procedure Result

101 ($\sim$ 6.3%) OTUs are discovered at $\alpha = 0.05$.



Three Statistical Challenges

- Overdispersion
- Heterogeneous library sizes
- Heterogeneous total feature counts

Overdispersion

OTU	y_{m1}	y_{m2}	$\cdots$	y_{m9}
1	4230	3563	$\cdots$	1954
2	3523	3222	$\cdots$	1559
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
1592	0	2	$\cdots$	0

$\mu(Y_{mn}) < Var(Y_{mn})$: 1451 (91.1%) OTUs
$\mu(Y_{mn}) = Var(Y_{mn})$: 10 (0.6%) OTUs
$\mu(Y_{mn}) > Var(Y_{mn})$: 131 (8.2%) OTUs

Heterogeneous Library Sizes

OTU	y_{m1}	y_{m2}	$\cdots$	y_{m9}
1	4230	3563	$\cdots$	1954
2	3523	3222	$\cdots$	1559
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
1592	0	2	$\cdots$	0
Total ($y_{.n}$)	81839	67861	$\cdots$	32073

Range of the library sizes: (32073, 92383)

Standard deviation: 17898

Problem for Heterogeneous Library Sizes

Feature	Sample 1	Sample 2
1	10	20
2	100	200
⋮	⋮	⋮
N	103	206
Total ($y_{.n}$)	2107	4214

Heterogenous Total Feature Counts

OTU	y_{m1}	y_{m2}	$\cdots$	y_{m9}	Total ($y_{m.}$)
1	4230	3563	$\cdots$	1954	33243
2	3523	3222	$\cdots$	1559	30809
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1592	0	2	$\cdots$	0	10

Range of the total feature counts: (10, 33243)

Standard deviation: 1663

Problem for Heterogeneous Total Feature Counts

$Y_{mn} \sim \text{Poisson}(\mu_{mn})$ and $\log(\mu_{mn}) = \alpha_m + \beta_m x_n$

OTU	$\hat{\beta}_m$	$y_{m.}$	$\widehat{IFDR}_m$	Discovered
1	-1.09	11	0.29	$\times$
2	0.19	911	0.003	✓

BH procedure provides similar results (Habiger, Watts, and Anderson, 2017).

Proposed New Methods

1. Address all the three statistical challenges:
   - Overdispersion
   - Heterogeneous library sizes
   - Heterogeneous total feature counts
2. Control the *FDR*

Proposed Model

Assume $Y_{mn} \sim \text{Poisson}(\mu_{mn})$ and

$$\log(\mu_{mn}) = \alpha_m + \beta_m x_n + d_n. \tag{4}$$

α_m: adjusts heterogeneous total feature counts effects

β_m: quantifies the association between the mean counts and the trait values

d_n: adjusts heterogeneous library sizes effects

Proposed Model

$Y_{mn} \sim \text{Poisson}(\mu_{mn})$ and $\log(\mu_{mn}) = \alpha_m + \beta_m x_n + d_n$

- pmf notation: $p(\boldsymbol{y}_m | \alpha_m, \beta_m, \boldsymbol{d})$

Assume $P(\beta_m = \gamma_k) = \pi_k$

Mixture probability mass function

$$p(\boldsymbol{y}_m | \alpha_m, \boldsymbol{\gamma}, \boldsymbol{d}, \boldsymbol{\pi}) = \pi_0 p(\boldsymbol{y}_m | \alpha_m, \gamma_0, \boldsymbol{d}) + \ldots \quad (5)$$

$$+ \pi_k p(\boldsymbol{y}_m | \alpha_m, \gamma_K, \boldsymbol{d}) \quad (6)$$

Null Hypothesis

The m^{th} null hypothesis can be specified as $H_m : \beta_m = 0$.

Alternatively, the m^{th} empirical null hypothesis is $H_m : \beta_m = \hat{\gamma}_0$.

Oracle Procedure

1. Compute *IFDR*

$$IFDR_m = P(\beta_m = \gamma_0 | \boldsymbol{y}_m; \alpha_m, \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{d}) \tag{7}$$

$$= \frac{\pi_0 p(\boldsymbol{y}_m | \alpha_m, \beta_m = \gamma_0, \boldsymbol{d})}{p(\boldsymbol{y}_m | \alpha_m, \boldsymbol{\gamma}, \boldsymbol{d}, \boldsymbol{\pi})}. \tag{8}$$

2. Reject j null hypotheses with smallest *IFDR*:

$$j = \max\{m : \sum_{i=1}^{m} IFDR_{(i)} \leq m\alpha\}$$

Parameter Estimation

Three ways to estimate $\alpha_m, \boldsymbol{\pi}, \boldsymbol{\gamma}, \boldsymbol{d}$:

1. Plug in the MLEs of $\alpha_m, \boldsymbol{\pi}, \boldsymbol{\gamma}$, and $\boldsymbol{d}$, which are obtained by the EM algorithm.
2. Condition on either library sizes $y_{.n}$ or total feature counts $y_{m.}$. Obtain the MLEs of the remaining parameters.
3. Condition on both library sizes $y_{.n}$ and total feature counts $y_{m.}$. Obtain the MLEs of $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$.

Analytical Assessment

- $Var(Y_{mn}) \geq E(Y_{mn})$
- For some $\beta_m \neq 0$,
  $IFDR_m = P(\beta_m = 0 | \boldsymbol{y}_m; \alpha_m, \boldsymbol{\pi}, \boldsymbol{\gamma}, \boldsymbol{d}) \rightarrow 1$ as $y_{m.} \rightarrow \infty$.
- FDR is controlled.

Aims for Simulation Study

1. Empirical assessment of the three statistical challenges
2. Comparison with other methods
3. Robustness check

Simulation Procedure

Recall that $Y_{mn} \sim \text{Poisson}(\mu_{mn})$ satisfying model

$$\log(\mu_{mn}) = \alpha_m + \beta_m x_n + d_n$$

with mixture density

$$p(\boldsymbol{y}_m | \alpha_{m.}, \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{d}) = \sum_{k=0}^{K} \pi_k p(\boldsymbol{y}_m | \alpha_{m.}, \beta_m = \gamma_k, \boldsymbol{d}).$$

1. $\alpha_m, d_n, \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{x}$
2. Simulate 10,000 or 1592 features

Methods for Comparison

Methods compared in literature:

edgeR, DESeq, NBPSeq, TSPM, baySeq, EBSeq, NOISeq, SAMSeq, ShrinkSeq, and two tests based on limma.

Methods that provide R packages:

QuasiSeq, PoissonSeq, BBSeq, BMDE, ShrinkBayes, and *clFDR*.

Assessment Metrics

- the average *FDR*
- the average power
- the average discovered effect size

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289–300.

Efron, Bradley (2008). "Microarrays, Empirical Bayes and the Two-Groups Model". In: *Statist. Sci.* 23.1, pp. 1–22.

Habiger, Joshua, David Watts, and Michael Anderson (2017). "Multiple Testing with Heterogeneous Multinomial Distributions". In: *Biometrics* 73.2, pp. 562–570.

z test statistic

Assume $Y_{mn} \sim \text{Poisson}(\mu_{mn})$ and $\log(\mu_{mn}) = \alpha_m + \beta_m x_n$.

$$z_m = \frac{\hat{\beta}_m}{I_\beta^{-1/2}} \tag{9}$$

The log likelihood of the Poisson log linear model is

$$l(\boldsymbol{\mu}, \boldsymbol{y}) = \sum (y_{mn} \log \mu_{mn} - \mu_{mn}) \tag{10}$$

The asymptotic variance of $\hat{\beta}$ is i_β^{-1}, where i_β is the negative matrix of second derivative of equation (10).

IFDR vs. *FDR*

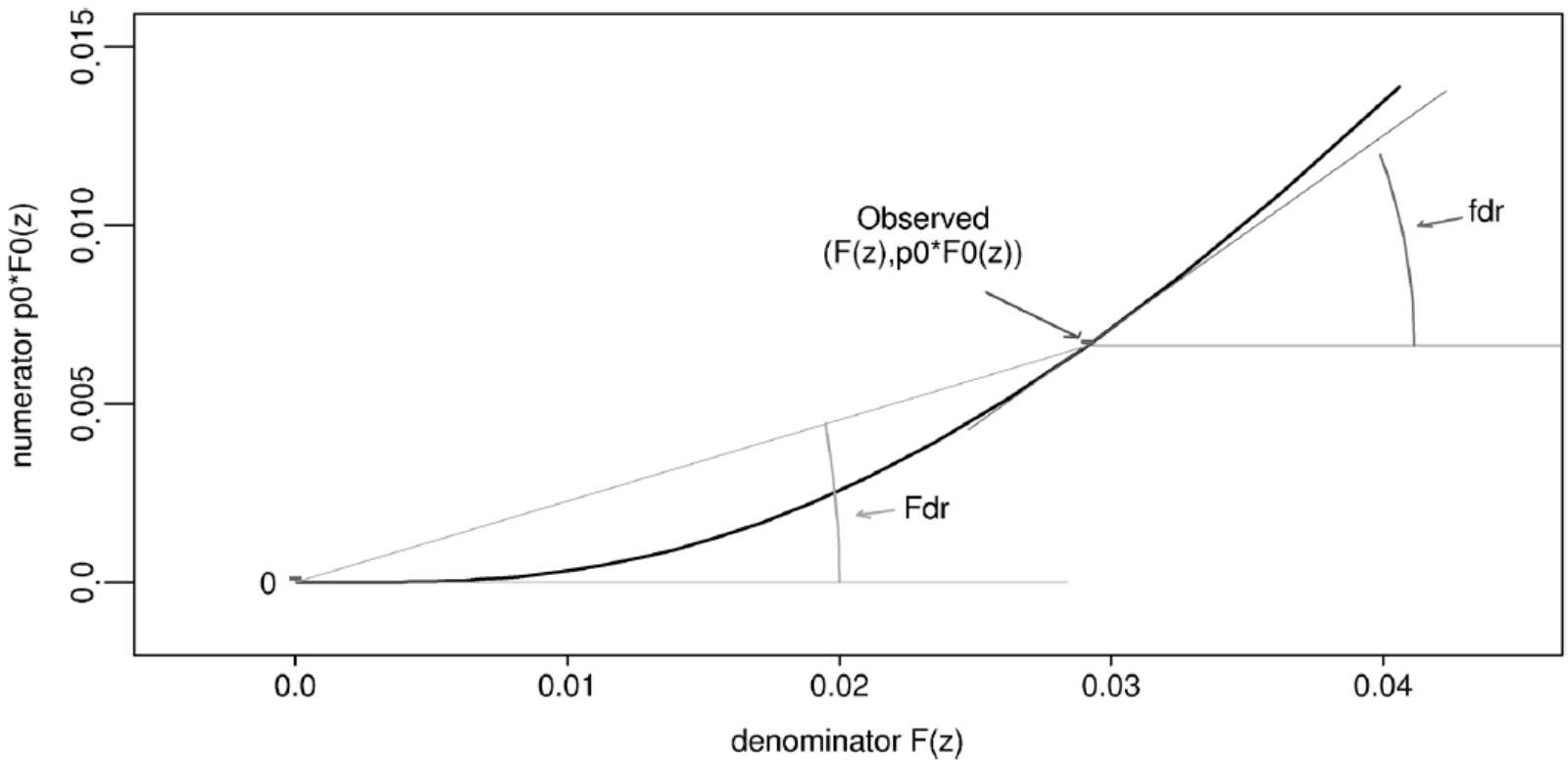


FIG. 2. *Relationship of Fdr(z) to fdr(z). Heavy curve plots numerator of Fdr, $p_0F_0(z)$, versus denominator $F(z)$; fdr(z) is slope of tangent, Fdr slope of secant.*

Negative Binomial Distribution

Hierarchical Poisson-gamma distribution
Assume $Y|\mu \sim \text{Poisson}(\mu)$ and $\mu \sim \text{gamma}(\alpha, \beta)$.

$$P(Y = y) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \int_0^{\infty} \frac{e^{-\lambda}\lambda^{y}}{y!} \lambda^{\alpha-1} e^{-\lambda/\beta} d\lambda \tag{11}$$

$$= \frac{1}{y!\Gamma(\alpha)\beta^{\alpha}} \int_0^{\infty} \lambda^{y+\alpha-1} e^{-\lambda(1+\frac{1}{\beta})} d\lambda \tag{12}$$

$$= \frac{1}{\Gamma(y+1)\Gamma(\alpha)\beta^{\alpha}} \Gamma(y+\alpha)(\frac{\beta}{\beta+1})^{y+\alpha} \tag{13}$$

$$= \binom{\alpha+y-1}{y} (\frac{1}{\beta+1})^{\alpha} (1 - \frac{1}{\beta+1})^{y} \tag{14}$$