Identifying relevant covariates in RNA-seq analysis by pseudo-variable augmentation

Yet Nguyen, Ph.D.[1] and Dan Nettleton, Ph.D.[2]

2024-11-02

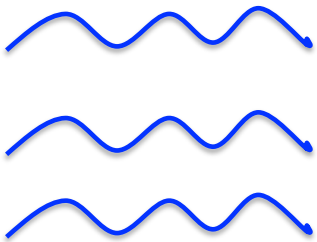
[1]Department of Mathematics and Statistics, Old Dominion University
[2]Department of Statistics, Iowa State University

Experimental Design

Pig

Blood Sample

RNA sample



Diet
Line
RFI

Basophil
Eosinophil
Lymphocyte
Monocyte
Neutrophil

Block
Order
RIN before globin depletion (GD)
RIN after GD
RNA Concentration before GD
RNA Concentration after GD

Prototypical RNA-seq Dataset

	Treatment 1				Treatment 2			
	u_{11}	u_{12}	$\cdots$	u_{1n_1}	u_{21}	u_{22}	$\cdots$	u_{2n_2}
$\boldsymbol{x}_1$	0	0	$\cdots$	0	1	1	$\cdots$	1
$\boldsymbol{x}_2$	0.5	0.95	$\cdots$	-1.42	-.45	.89	$\cdots$	1.2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\boldsymbol{x}_k$	0	1	$\cdots$	0	1	0	$\cdots$	0
gene 1	10	13	$\cdots$	2017	31	975	$\cdots$	3289
gene 2	0	2	$\cdots$	1	0	0	$\cdots$	1
gene 3	1	3	$\cdots$	0	0	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
gene G	17301	2464	$\cdots$	7345	3214	534	$\cdots$	934

How to Handle the Available Covariates?

- Including all available covariates (Full)
- Excluding all available covariates (OnlyLine)
- Backward selection that maximizes the number of DE genes with respect to the main factor of interest (BS15, Nguyen et al. 2015)

Our Proposed Method (Nguyen and Nettleton 2024+)

- Using `limma-voom` (Law et al. (2014)) for differential expression analysis to obtain vectors of p-values of tests for significance of regression coefficients w.r.t each of the covariates
- Selecting the most relevant covariates by a backward selection strategy intending to control the false seletion rate (FSR) using pseudo-variables (Wu et al. (2007), 'Controlling Variable Selection by the Addition of Pseudovariables', JASA)

- Wu et al. (2007) method published for one response variable

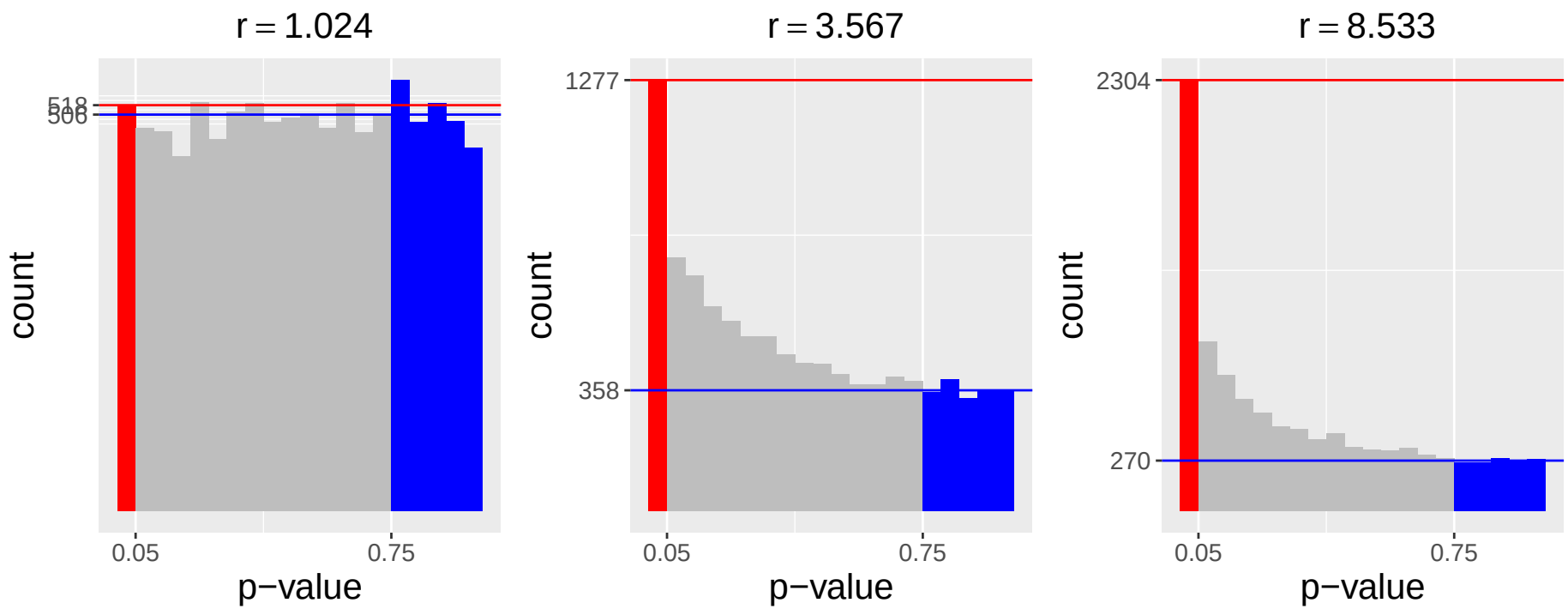
- We extend Wu et al. (2007)'s method to thousands of response variables

Measure of Covariate Relevance

Definition

With $\mathbb{1}$ representing an indicator function, a relevance measure for covariate j is defined as

$$r(\boldsymbol{p}_j) = \frac{\sum_{g=1}^{G} \mathbb{1}(p_{gj} \leq 0.05)}{\max\{\sum_{g=1}^{G} \mathbb{1}(p_{gj} \geq 0.75)/5, 1\}}. \quad (1)$$



Backward Selection to Control FSR

- Run backward selection procedure using $r(\cdot)$ on k_T covariates of $\boldsymbol{X}$
- Let $BS(\boldsymbol{X}, \lambda)$ denote the subset of $\boldsymbol{X}$ selected by this backward selection, i.e., the largest subset of $\boldsymbol{X}$ for which each variable has r-value at least λ
- Define $S(\lambda) = \text{Card}\{BS(\boldsymbol{X}, \lambda)\}$. Then $S(\lambda) = R(\lambda) + I(\lambda)$, where $R(\lambda)$ and $I(\lambda)$ denote the number of selected relevant and irrelevant covariates, respectively
- False selection rate (FSR) is calculated as $\alpha(\lambda) = \frac{E(I(\lambda))}{E(S(\lambda)+1)}$
- Calculate the tuning parameter λ_* to control FSR at level α_0

$$\lambda_* = \inf\{\lambda : \alpha(\lambda) \leq \alpha_0\}.$$

Estimating FSR I

Generate B sets of k_P pseudo-variables $\boldsymbol{Z}_b$

Define $\alpha_P(\lambda) = \frac{E(I^*_{P,b}(\lambda))}{E(1+S_{P,b}(\lambda))}$ where

- $R_{P,b}(\lambda)$: number of truly relevant covariates selected from $\boldsymbol{X}, \boldsymbol{Z}_p$
- $I_{P,b}(\lambda)$: number of truly irrelevant covariates selected from $\boldsymbol{X}, \boldsymbol{Z}_p$
- $I^*_{P,b}(\lambda)$: number of pseudo-covariates selected from $\boldsymbol{X}, \boldsymbol{Z}_p$
- $S_{P,b} = R_{P,b}(\lambda) + I_{P,b}(\lambda) + I^*_{P,b}(\lambda)$

Estimating FSR II

Assumptions

(A1) $E(I(\lambda)) = E(I_{P,b}(\lambda)) = k_U E(I^*_{P,b}(\lambda))/k_P$, where k_U is the unknown number of truly irrelevant covariates

(A2) $E(R_{P,b}(\lambda)) = E(R(\lambda))$

- (A1) & (A2) imply: $\alpha_P(\lambda) = \frac{k_P \alpha(\lambda)}{k_P \alpha(\lambda) + k_U}$
- Let $\bar{I}^*_P(\lambda) = B^{-1} \sum_{b=1}^{B} I^*_{P,b}(\lambda)$, $\bar{S}_P(\lambda) = B^{-1} \sum_{b=1}^{B} S_{P,b}(\lambda)$
- Estimate $\alpha_P(\lambda)$ by $\widehat{\alpha}_P(\lambda) = \frac{\bar{I}^*_P(\lambda)}{1 + \bar{S}_P(\lambda)}$
- If k_U is known, estimate $\alpha(\lambda)$ by solving

$$\widehat{\alpha}_P(\lambda) = \frac{k_P \alpha(\lambda)}{k_P \alpha(\lambda) + k_U}$$

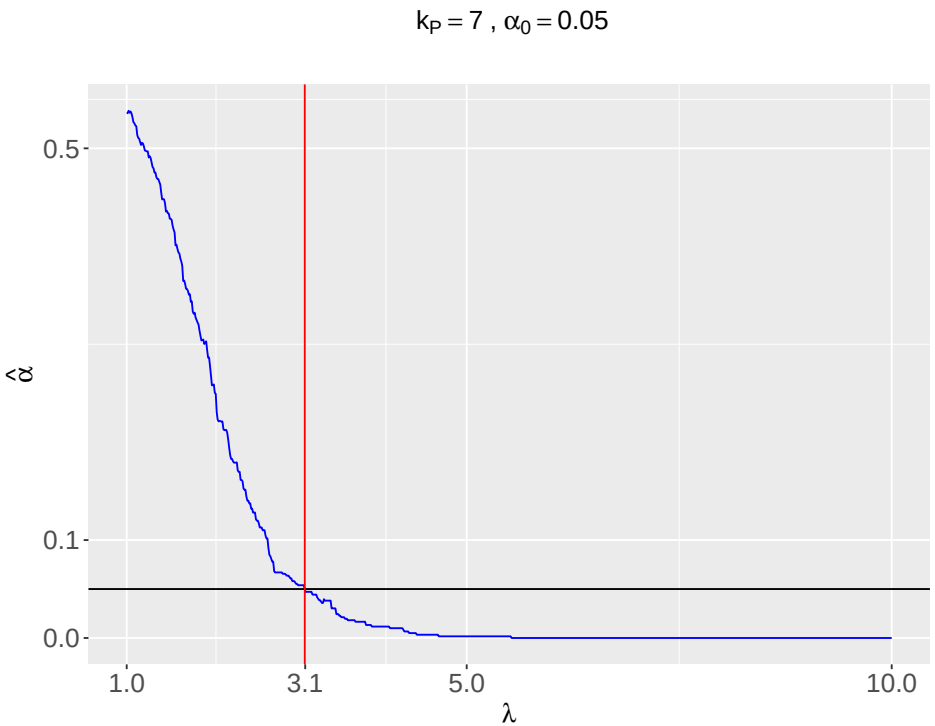
Generating Pseudo-covariates $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{k_P})$

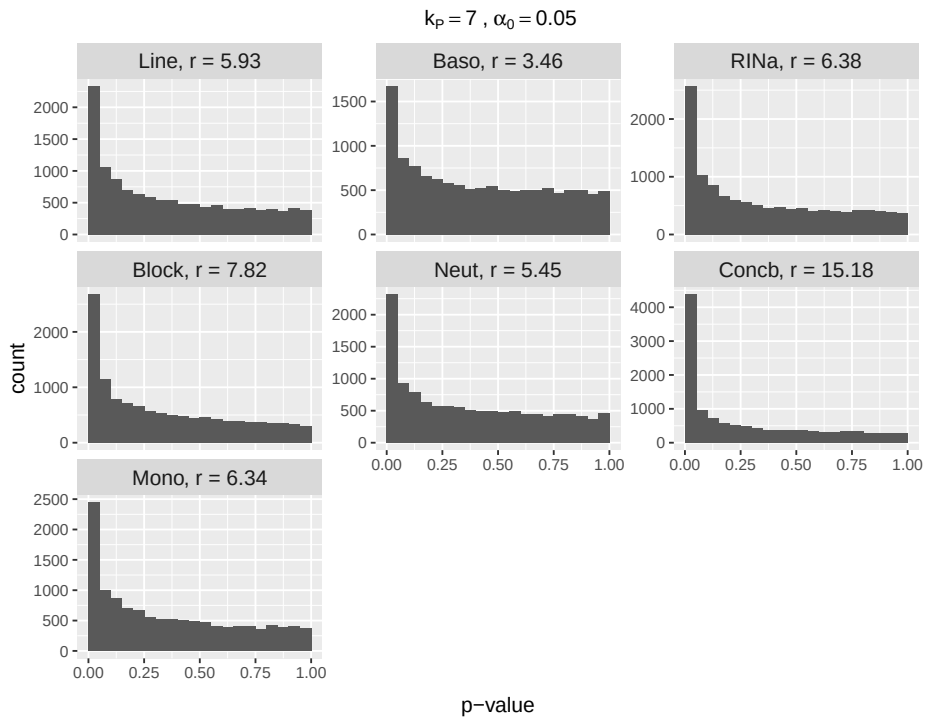
- Option 1 (WN): $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{k_P}$ i.i.d. $\sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{1})$
- Option 2 (RX): The n rows of $\boldsymbol{Z}$ are obtained by randomly permuting the rows and the columns of $\boldsymbol{X}$
- Options 3 & 4 (OWN & ORX): $(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}})\boldsymbol{Z}$, where $\boldsymbol{H}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, where $\boldsymbol{Z}$ is generated either by option 1 or 2, respectively

RFI RNA-seq Data Analysis

Table 1: Covariates removed from the full model and their r values at each iteration of the backward selection algorithm applied to the RFI RNA-seq dataset.

Iteration	1	2	3	4	5	6	7	8	9	10	11	12	13
Covariate	RINb	Eosi	Order	Conca	Diet	RFI	Lymp	Baso	RINa	Block	Neut	Concb	Mono
r	0.26	0.49	0.62	0.65	0.53	2.07	2.87	3.46	6.3	7.71	7.85	9.42	11.45





Simulation Study - Setting

Table 2: Six simulation scenarios corresponding to six sets of truly relevant covariates.

Number of relevant covariates k_R	Relevant covariates
0	
1	Mono
2	Concb, Mono
6	Baso, RINa, Block, Neut, Concb, Mono
7	Lymp, Baso, RINa, Block, Neut, Concb, Mono
8	RFI, Lymp, Baso, RINa, Block, Neut, Concb, Mono

- Number of genes: 2000
- Number of replications: 100

Simulation Study - FSR Results

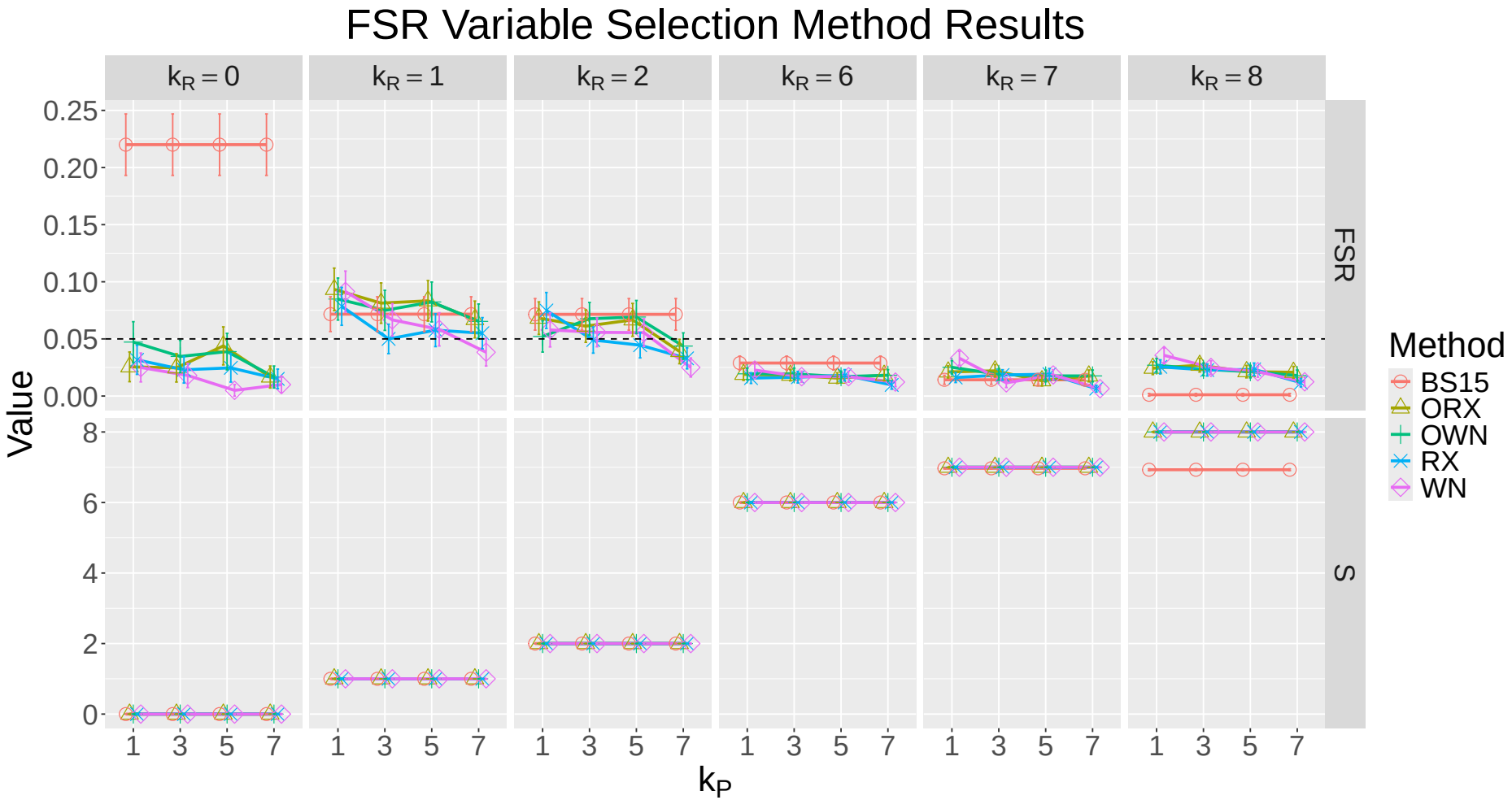


Figure 1: The figure displays the variable selection performance of four variants of the proposed method and BS15.

Simulation Study - Differential Expression Analysis Results

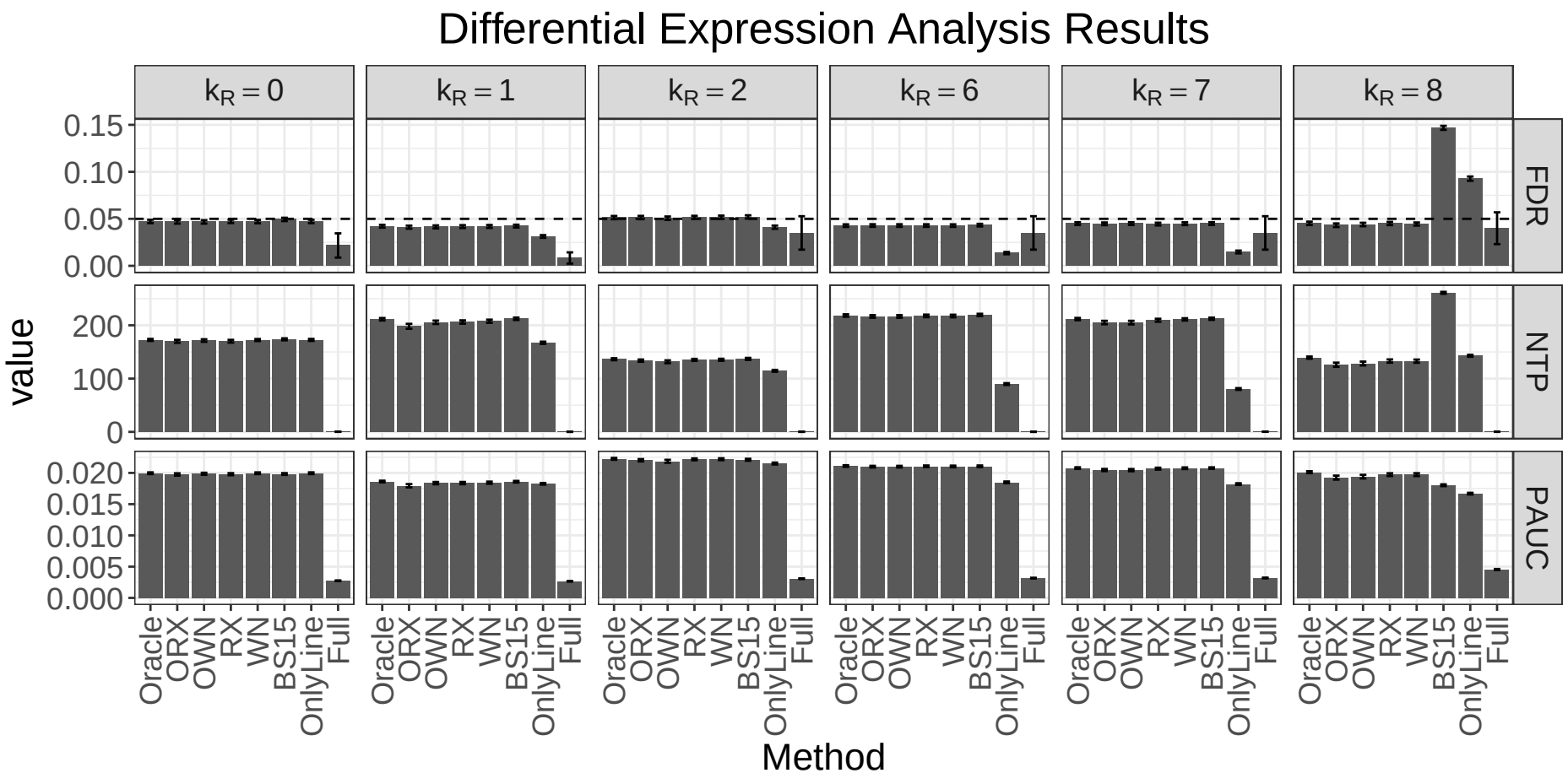


Figure 2: The figure presents the performance of differential expression analysis of the twelve methods.

Conclusion

- The proposed covariate selection method control FSR well
- The selected model has good performance in identifying DE genes in terms of
  - FDR control
  - Ability to distinguish EE genes and DE genes
- The proposed method is available at github.com/ntyet/csrnaseq
- Contact: Yet Nguyen, ynguyen@odu.edu

Thank you!

References I

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014), "Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts," *Genome Biology*, 15, R29. https://doi.org/10.1186/gb-2014-15-2-r29.

Nguyen, Y., and Nettleton, D. (2024+), "Identifying relevant covariates in RNA-seq analysis by pseudo-variable augmentation," *Journal of Agricultural, Biological and Environmental Statistics*, accepted.

Nguyen, Y., Nettleton, D., Liu, H., and Tuggle, C. K. (2015), "Detecting differentially expressed genes with RNA-seq data using backward selection to account for the effects of relevant covariates," *Journal of Agricultural, Biological, and Environmental Statistics*, 20, 577–597. https://doi.org/10.1007/s13253-015-0226-1.

Wu, Y., Boos, D. D., and Stefanski, L. A. (2007), "Controlling variable selection by the addition of pseudovariables," *Journal of the American Statistical Association*, Taylor & Francis, 102, 235–243. https://doi.org/10.1198/016214506000000843.