



Statistical Rigor in Genomics Data Analysis

Jingyi Jessica Li

Department of Statistics
University of California, Los Angeles

<http://jsb.ucla.edu>

Statistical rigor challenges in genomics data analysis

What is statistical rigor?

- Performance guarantee of statistical methods, e.g.,
 - { **p-values**: uniformly distributed between 0 and 1 under the nulls
 - { **confidence intervals**: coverage probabilities = the claimed level (e.g., 95%)
 - { **false discovery rate (FDR)**: average ($\#$ false discoveries)/($\#$ discoveries) in permutation analysis = the claimed level (e.g., 5%)



Statistical rigor challenges in genomics data analysis

What is statistical rigor?

- Performance guarantee of statistical methods, e.g.,
 - { **p-values**: uniformly distributed between 0 and 1 under the nulls
 - { **confidence intervals**: coverage probabilities the claimed level (e.g., 95%)
 - { **false discovery rate (FDR)**: average ($\#$ false discoveries)/($\#$ discoveries) in **permutation analysis** the claimed level (e.g., 5%)

Why is statistical rigor challenging in genomics data analysis?

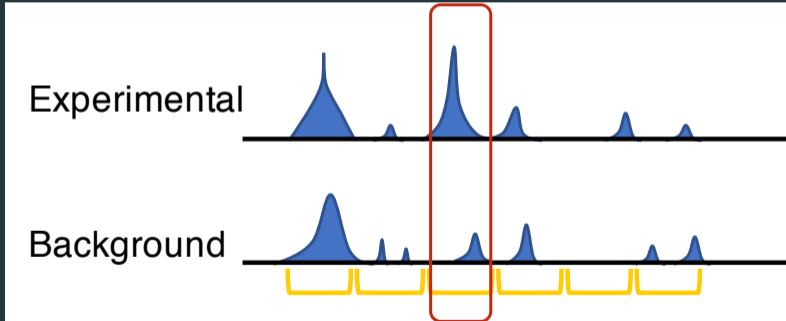
- **New** and **complex** data types
- **Fast** method development



Common causes of invalid p-values

1. Mis-formulation of a **two-sample test** as a one-sample test

- Peak calling from ChIP-seq data



Common causes of invalid p-values

1. Mis-formulation of a **two-sample test** as a one-sample test

- Peak calling from CHIP-seq data (e.g., MACS and HOMER)

	a region	background count	experimental count
{	random variable (hypothetical)	X	Y
	random observation (data)	x	y



Common causes of invalid p-values

1. Mis-formulation of a **two-sample test** as a one-sample test

- Peak calling from ChIP-seq data (e.g., MACS and HOMER)

a region	background count	experimental count
{ random variable (hypothetical)	X	Y
random observation (data)	x	y
{ p-value = $\mathbb{P}(Y \geq y)$ where $Y \sim \text{Poisson}(x)$		correct?



Common causes of invalid p-values

1. Mis-formulation of a **two-sample test** as a one-sample test

- Peak calling from ChIP-seq data (e.g., MACS and HOMER)

a region	background count	experimental count
{ random variable (hypothetical)	X	Y
{ random observation (data)	x	y
{ p-value = $\mathbb{P}(Y = y)$ where $Y \sim \text{Poisson}(x)$ correct?		
{ No , because it assumes $Y \sim \text{Poisson}(\)$ and tests		

$$H_0 : \mu = x \quad \text{vs.} \quad H_1 : \mu > x ;$$

which treats x as a fixed parameter and ignores its randomness



Common causes of invalid p-values

1. Mis-formulation of a **two-sample test** as a one-sample test

- How to perform a two-sample test when the sample size is 1 vs. 1?
 - { p-value calculation is difficult...



Common causes of invalid p-values

1. Mis-formulation of a **two-sample test** as a one-sample test

- How to perform a two-sample test when the sample size is 1 vs. 1?
 - { p-value calculation is difficult...
 - { but, p-values are just intermediates for FDR control in high-throughput data analysis



Our solution: Clipper

1. Mis-formulation of a **two-sample test** as a one-sample test

- How to perform a two-sample test when the sample size is 1 vs. 1?
 - { p-value calculation is difficult...
 - { but, p-values are just intermediates for FDR control in high-throughput data analysis

Clipper: p-value-free FDR control on high-throughput data from two conditions

 Xinzhou Ge,  Yiling Elaine Chen,  Dongyuan Song, MeiLu McDermott, Kyla Woysner,
 Antigoni Manousopoulou,  Ning Wang,  Wei Li,  Leo D.Wang,  Jingyi Jessica Li

doi: <https://doi.org/10.1101/2020.11.19.390773>

| accepted by *Genome Biology*



Our solution: Clipper

- ^ Does not

- { require high-resolution p-values
- { assume parametric distributions
- { require many replicates

- ^ Two components

- { contrast scores
- { α

sample size n vs. n
 d features
FDR threshold q

- ^ Applications

- { ChIP-seq
- { mass spectrometry
- { bulk and single-cell RNA-seq
- { Hi-C

Our solution: Clipper

- ^ Peak calling from ChIP-seq data
 - { as an add-on, **Clipper** improves the FDR control of MACS2 and HOMER

Common causes of invalid p-values

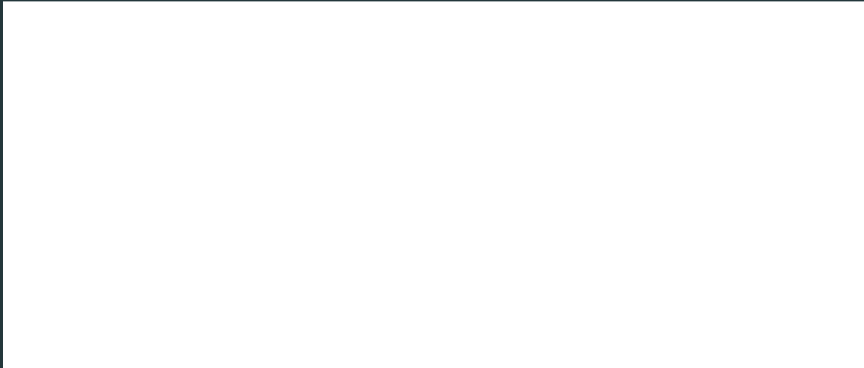
2. Mis-specification of a **parametric model** that does not fit data well

^ Identification of differentially expressed (DE) genes from RNA-seq data



Common causes of invalid p-values

2. Mis-specification of a **parametric model** that does not fit data well
 - ^ Identification of differentially expressed genes (DEGs) from RNA-seq data
 - { **FDR Check**: permute samples between conditions (no true DEGs)



Our recommendation: Mann-Whitney-Wilcoxon rank-sum test

2. Mis-specification of a **parametric model** that does not fit data well

{ **Recommendation**: consider non-parametric methods when sample size is large: $(\text{sample size})/(\text{\# number of parameters}) \geq 20$



| collaboration with Dr. Yumei Li in Dr. Wei Li's lab (UC Irvine)

Common causes of invalid p-values

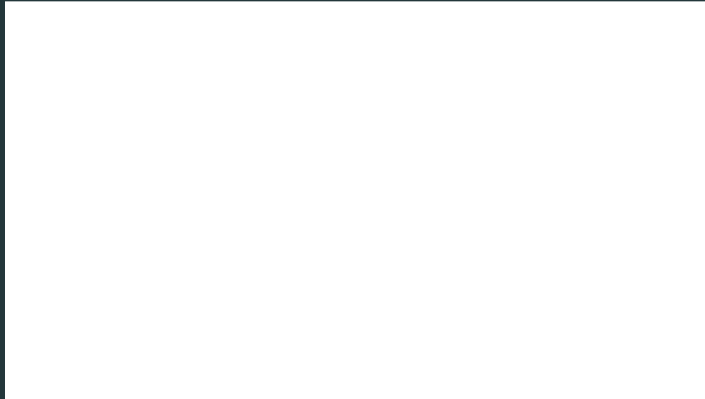
3. Mis-treatment of inferred covariates as observed

- ^ Identification of DEGs along cell pseudotime from scRNA-seq data
 - { **Pseudotime**: a latent "temporal" variable that reflects a cell's relative transcriptome status among all cells
 - { **Pseudotime inference** (trajectory inference): estimate the pseudotime of cells, i.e., order cells along a trajectory based on transcriptome similarities
 - { Popular methods:
 - Monocle3 (Trapnell et al. 2014)
 - TSCAN (Ji et al. 2016)
 - Slingshot (Street et al. 2018)

Common causes of invalid p-values

3. Mis-treatment of inferred covariates as observed

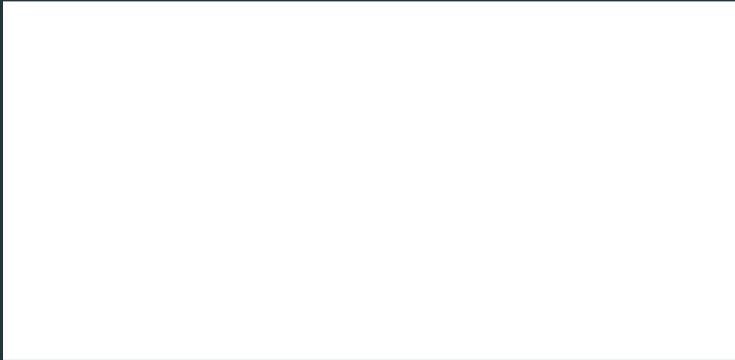
- ^ Identification of DEGs along cell pseudotime from scRNA-seq data



Common causes of invalid p-values

3. Mis-treatment of inferred covariates as observed

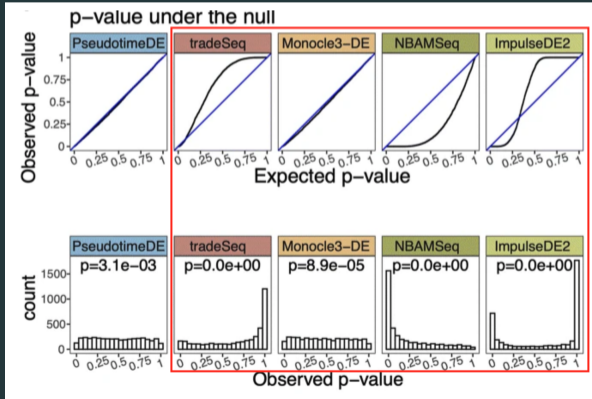
- Identification of DEGs along cell pseudotime from scRNA-seq data
 - Cell pseudotime is **random**



Common causes of invalid p-values

3. Mis-treatment of inferred covariates as observed

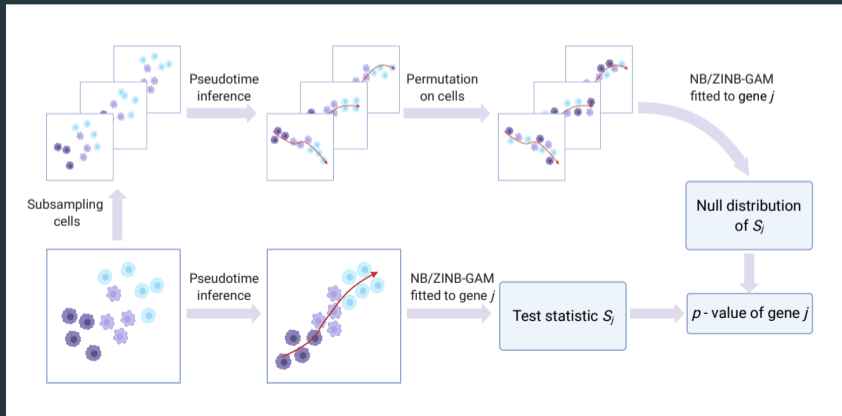
- Identification of DEGs along cell pseudotime from scRNA-seq data
 - Existing methods treat cell pseudotime as a **observed covariate**



Our solution: PseudotimeDE

3. Mis-treatment of inferred covariates as observed

- Identification of DEGs along cell pseudotime from scRNA-seq data
 - **PseudotimeDE** considers the uncertainty of pseudotime inference



Our solution: PseudotimeDE

3. Mis-treatment of inferred covariates as observed

- Identification of DEGs along cell pseudotime from scRNA-seq data
 - **PseudotimeDE** generates well-calibrated p-values for FDR control & uses a generalized additive model for good power

Method | [Open Access](#) | [Published: 29 April 2021](#)

PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p -values from single-cell RNA sequencing data

[Dongyuan Song](#) & [Jingyi Jessica Li](#) 

[Genome Biology](#) **22**, Article number: 124 (2021) | [Cite this article](#)

3221 Accesses | **1** Citations | **32** Altmetric | [Metrics](#)



Summary

Three common causes of invalid p-values in genomics data analysis

1. Mis-formulation of a two-sample test as a one-sample test
2. Mis-specification of a parametric model that does not fit data well
3. Mis-treatment of inferred covariates as observed



Summary

Three common causes of invalid p-values in genomics data analysis

1. Mis-formulation of a two-sample test as a one-sample test
2. Mis-specification of a parametric model that does not fit data well
3. Mis-treatment of inferred covariates as observed

Our proposals

1. **Clipper**: a p-value-free FDR control framework
2. **Renaissance** of **classical non-parametric methods** (e.g., Mann-Whitney-Wilcoxon rank-sum test) when sample sizes are large
3. **PseudotimeDE**: a method that identifies DEGs along cell pseudotime by considering pseudotime inference uncertainty



Patterns



Perspective

Statistical Hypothesis Testing versus Machine Learning Binary Classification: Distinctions and Guidelines

Jingyi Jessica Li^{1,*} and Xin Tong²

¹Department of Statistics, University of California, Los Angeles, CA 90095-1554, USA

²Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA

*Correspondence: jli@stat.ucla.edu

<https://doi.org/10.1016/j.patter.2020.100115>



