Statistical Rigor in Genomics Data Science

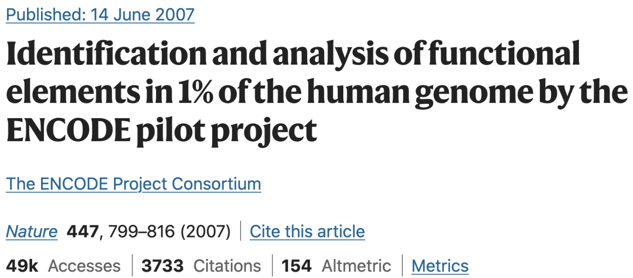
in honor of Peter Bickel's 82nd birthday

Jingyi Jessica Li

Associate Professor
Junction of Statistics and Biology (`http://jsb.ucla.edu`)
Department of Statistics
University of California, Los Angeles

My first encounter of genomics bioinformatics (2007)

Volume 447 Issue 7146, 14 June 2007



Published: 14 June 2007

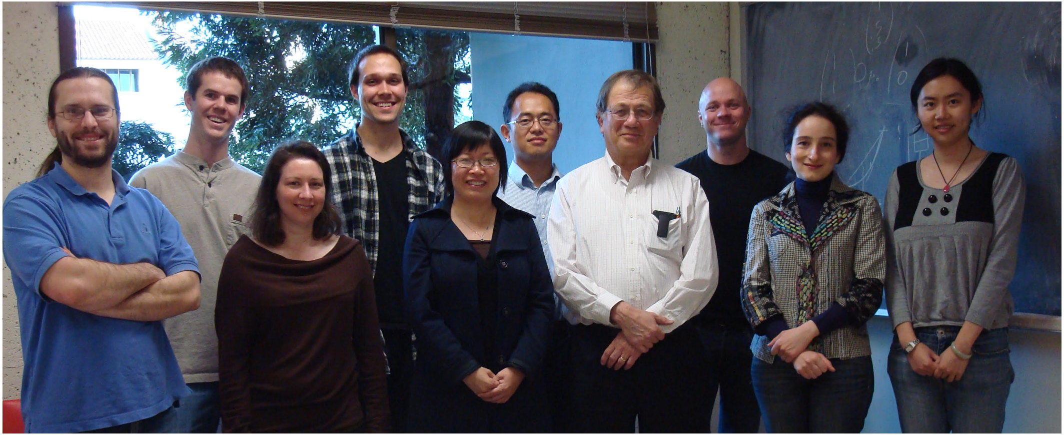
Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project

The ENCODE Project Consortium

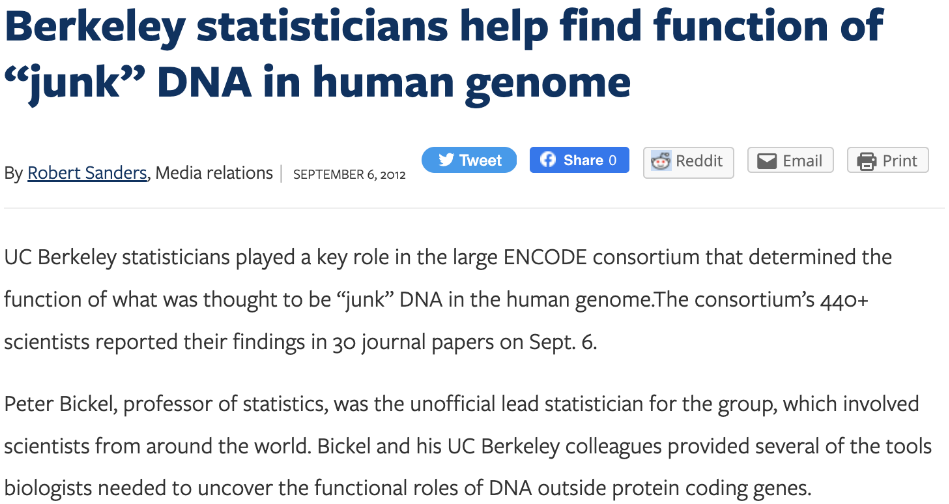
Nature **447**, 799–816 (2007) | Cite this article

49k Accesses | **3733** Citations | **154** Altmetric | Metrics

Peter and Haiyan’s ENCODE team (2007–2013)



Our ENCODE fruits (2012–2014)



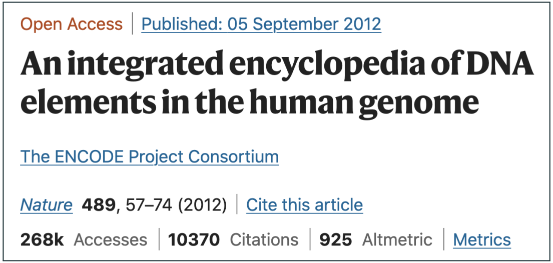
Berkeley statisticians help find function of “junk” DNA in human genome

By Robert Sanders, Media relations | SEPTEMBER 6, 2012

UC Berkeley statisticians played a key role in the large ENCODE consortium that determined the function of what was thought to be “junk” DNA in the human genome.The consortium’s 440+ scientists reported their findings in 30 journal papers on Sept. 6.

Peter Bickel, professor of statistics, was the unofficial lead statistician for the group, which involved scientists from around the world. Bickel and his UC Berkeley colleagues provided several of the tools biologists needed to uncover the functional roles of DNA outside protein coding genes.

Our ENCODE fruits (2012–2014)



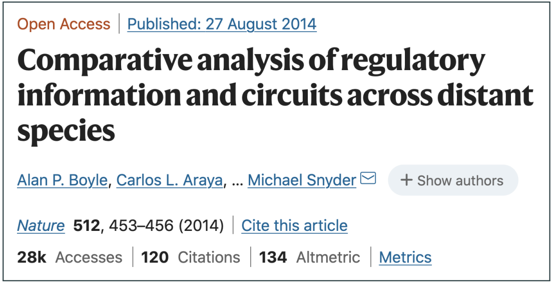
Open Access | Published: 05 September 2012

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium

Nature **489**, 57–74 (2012) | Cite this article

268k Accesses | **10370** Citations | **925** Altmetric | Metrics



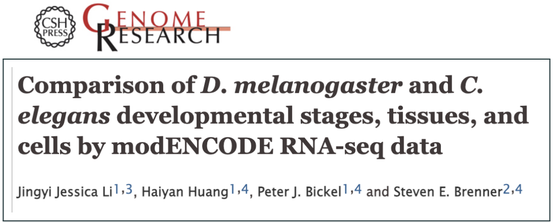
Open Access | Published: 27 August 2014

Comparative analysis of regulatory information and circuits across distant species

Alan P. Boyle, Carlos L. Araya, … Michael Snyder ✉ + Show authors

Nature **512**, 453–456 (2014) | Cite this article

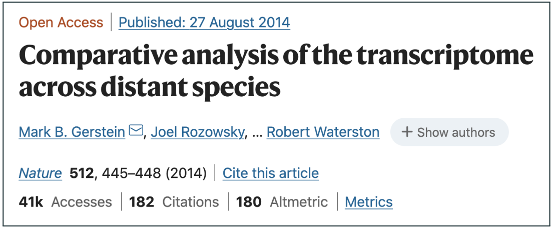
28k Accesses | **120** Citations | **134** Altmetric | Metrics



CSH PRESS GENOME RESEARCH

Comparison of *D. melanogaster* and *C. elegans* developmental stages, tissues, and cells by modENCODE RNA-seq data

Jingyi Jessica Li[1,3], Haiyan Huang[1,4], Peter J. Bickel[1,4] and Steven E. Brenner[2,4]



Open Access | Published: 27 August 2014

Comparative analysis of the transcriptome across distant species

Mark B. Gerstein ✉, Joel Rozowsky, … Robert Waterston + Show authors

Nature **512**, 445–448 (2014) | Cite this article

41k Accesses | **182** Citations | **180** Altmetric | Metrics

Some questions I had about bioinformatics methods

1. Are p-values valid?

2. Why not classical statistical methods?

3. What is the proper null hypothesis?

Statistical rigor in multiple testing

Criteria need calibration

- p-values $\sim$ (super-)uniform$[0,1]$ under the null hypotheses
- false discovery rate (FDR) $= \mathbb{E}\left[\frac{\#\text{ false discoveries}}{\#\text{ discoveries} \vee 1}\right] \leq$ the claimed level (e.g., 5%)

Statistical rigor in multiple testing

Criteria need calibration

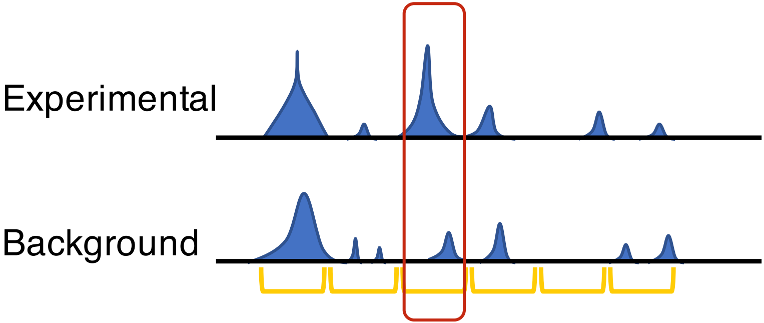
- p-values $\sim$ (super-)uniform$[0,1]$ under the null hypotheses
- false discovery rate (FDR) $= \mathbb{E}\left[\frac{\#\text{ false discoveries}}{\#\text{ discoveries} \vee 1}\right] \leq$ the claimed level (e.g., 5%)

Three common causes of ill-posed p-values

1. Formulation of a two-sample test as a one-sample test
2. Specification of a parametric model that does not fit data well
3. Treatment of inferred covariates as observed

1. Formulation of a two-sample test as a one-sample test

Example: peak calling from ChIP-seq data



1. Formulation of a two-sample test as a one-sample test

Peak calling from ChIP-seq data

- Popular software:
  - MACS [Zhang *et al.*, *Genome Biol*, 2008]; cited $>$ 10K times
  - HOMER [Heinz *et al.*, *Mol Cell*, 2010]; cited $\sim$ 8K times

- Formulation:

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?

1. Formulation of a two-sample test as a one-sample test

Peak calling from ChIP-seq data

- Formulation:

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?

  - No, because it assumes $Y \sim \text{Poisson}(\lambda)$ and tests

$$H_0 : \lambda = x \quad \text{vs.} \quad H_1 : \lambda > x \,,$$

which treats x as a fixed parameter and ignores its randomness

1. 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 ...

1. 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?

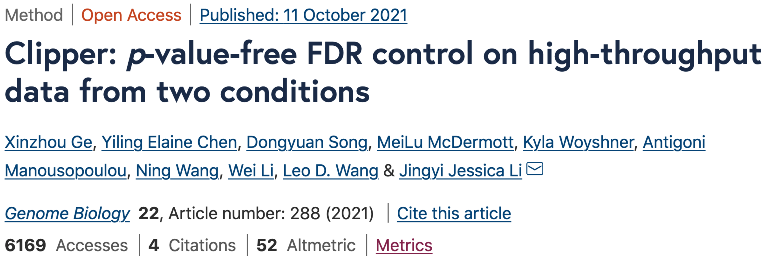
- p-value calculation is difficult ...
- but, p-values are just intermediates for FDR control in large-scale multiple testing

1. 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 large-scale multiple testing

Our solution: inspired by knockoffs [Barber and Candès, *Ann Stat*, 2015] (to be elaborated)



Method | Open Access | Published: 11 October 2021

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

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

Genome Biology **22**, Article number: 288 (2021) | Cite this article

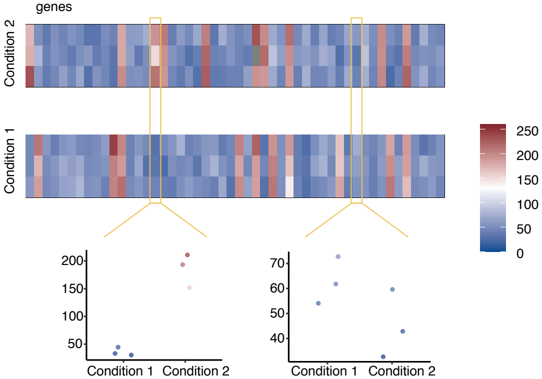
6169 Accesses | **4** Citations | **52** Altmetric | Metrics

2. Specification of a parametric model that does not fit data well

Example: identifying differentially expressed genes (DEGs) from RNA-seq data

- Popular software (originally designed for small sample sizes):
  - `edgeR` [Robinson *et al.*, *Bioinformatics*, 2014]; cited $\sim$ 24K times
  - `DESeq2` [Love *et al.*, *Genome Biol*, 2014]; cited $>$ 33K times

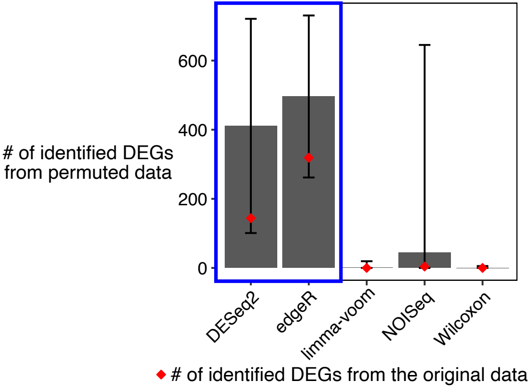
both assume a negative binomial distribution per gene and condition



2. Specification of a parametric model that does not fit data well

Identifying differentially expressed genes (DEGs) from RNA-seq data

- Check of false discoveries: permute individuals between conditions (no true DEGs)



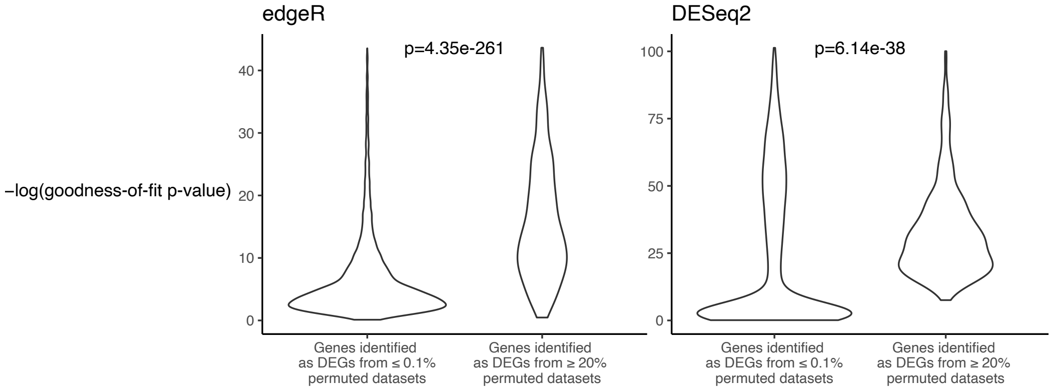
51 pre-nivolumab and 58 on-nivolumab anti-PD-1 therapy patients [Riaz *et al.*, *Cell*, 2017]

[Li et al., *Genome Biology*, 2022]

2. Specification of a parametric model that does not fit data well

Identifying differentially expressed genes (DEGs) from RNA-seq data

- Poor fit of negative binomial model ⟷ false positive DEGs



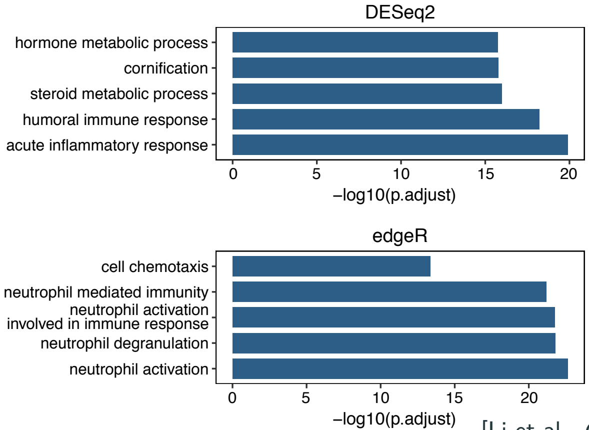
51 pre-nivolumab and 58 on-nivolumab anti-PD-1 therapy patients [Riaz *et al.*, *Cell*, 2017]

[Li et al., *Genome Biology*, 2022]

2. Specification of a parametric model that does not fit data well

Identifying differentially expressed genes (DEGs) from RNA-seq data

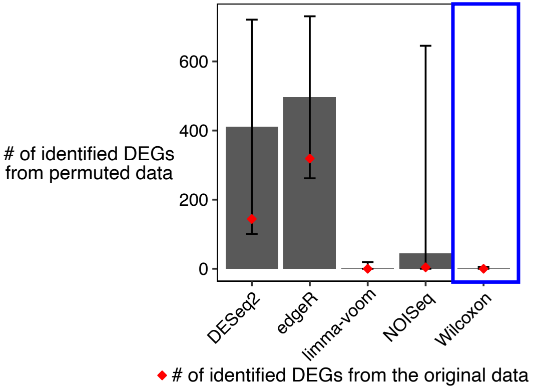
- False discoveries may mislead scientific conclusions



[Li et al., *Genome Biology*, 2022]

2. Specification of a parametric model that does not fit data well

Method choice: popular bioinformatics tools vs. general statistical methods?

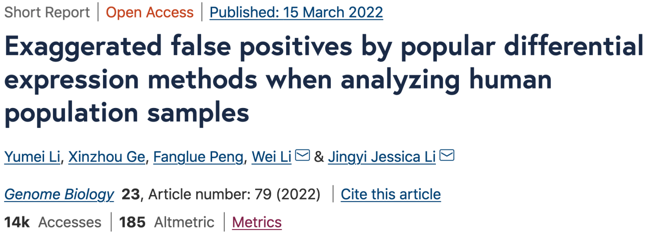


2. Specification of a parametric model that does not fit data well

Method choice: popular bioinformatics tools vs. general statistical methods?

Our recommendations for large-sample-sized data:

- sanity check: permutation
- consider non-parametric tests (e.g., Wilcoxon rank-sum test)



Short Report | Open Access | Published: 15 March 2022

Exaggerated false positives by popular differential expression methods when analyzing human population samples

Yumei Li, Xinzhou Ge, Fanglue Peng, Wei Li & Jingyi Jessica Li

Genome Biology **23**, Article number: 79 (2022) | Cite this article

14k Accesses | **185** Altmetric | Metrics

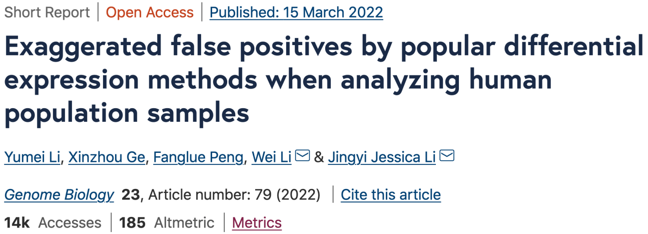
— collaboration with Dr. Yumei Li in Dr. Wei Li's lab (UC Irvine)

2. Specification of a parametric model that does not fit data well

Method choice: popular bioinformatics tools vs. general statistical methods?

Our recommendations for large-sample-sized data:

- sanity check: permutation
- consider non-parametric tests (e.g., Wilcoxon rank-sum test)



Short Report | Open Access | Published: 15 March 2022

Exaggerated false positives by popular differential expression methods when analyzing human population samples

Yumei Li, Xinzhou Ge, Fanglue Peng, Wei Li ✉ & Jingyi Jessica Li ✉

Genome Biology **23**, Article number: 79 (2022) | Cite this article

14k Accesses | **185** Altmetric | Metrics

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

- **What if sample sizes are small?**

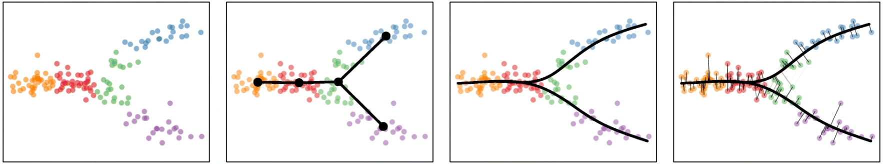
  Clipper is a non-parametric option (to be elaborated)

3. Treatment of inferred covariates as observed

Example: identifying DEGs along pseudotime from single-cell RNA-seq data

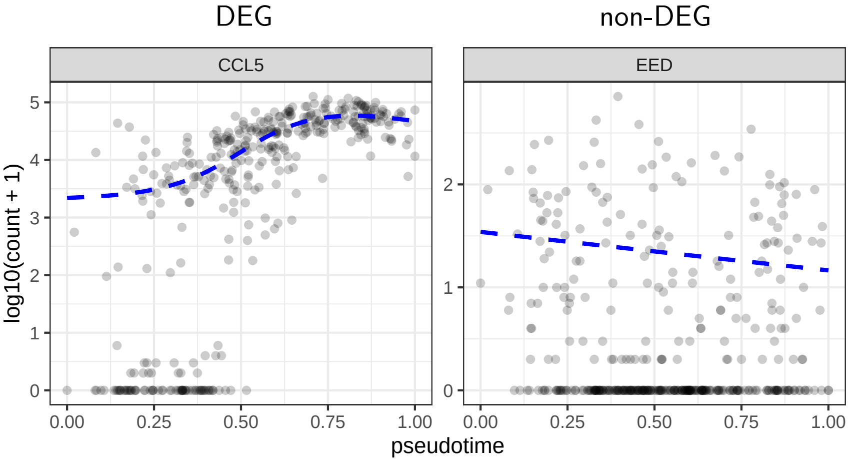
- Cell pseudotime: a latent "temporal" variable that reflects a cell's relative status among all cells
- Pseudotime inference: estimate the pseudotime of cells, i.e., order cells along a trajectory based on cells' high-dimensional gene expression vectors

- Popular software:
  - `Monocle3` [Trapnell *et al.*, *Nat Biotechnol*, 2014]; cited $>$ 2.8K times
  - `Slingshot` [Street *et al.*, *BMC Bioinform*, 2018]; cited 700 times



3. Treatment of inferred covariates as observed

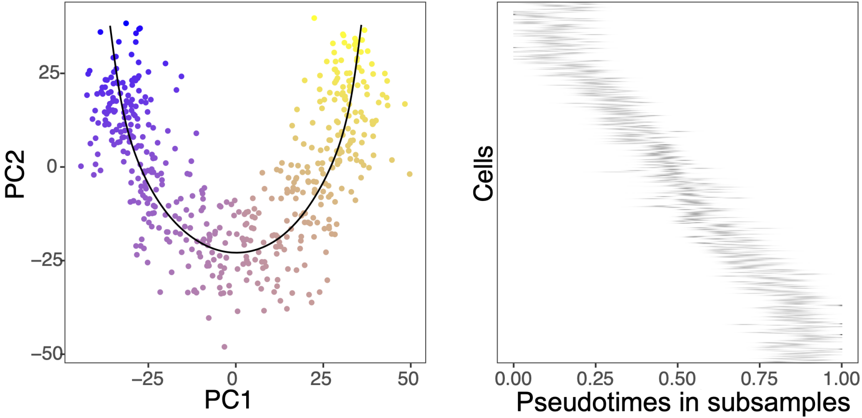
Identifying DEGs along inferred pseudotime from single-cell RNA-seq data



3. Treatment of inferred covariates as observed

Identifying DEGs along inferred pseudotime from single-cell RNA-seq data

- Cell pseudotime is inferred from the same data and thus random



3. Treatment of inferred covariates as observed

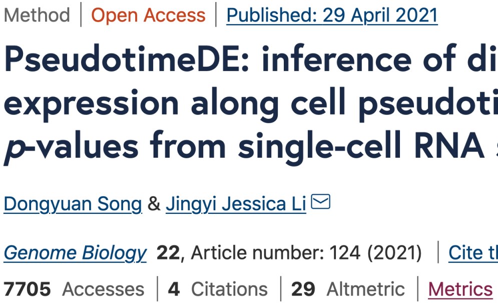
Identifying DEGs along inferred pseudotime from single-cell RNA-seq data

- However, existing methods treat cell pseudotime as an observed covariate

3. Treatment of inferred covariates as observed

Identifying DEGs along inferred pseudotime from single-cell RNA-seq data

- However, existing methods treat cell pseudotime as an observed covariate
- Our solution: PseudotimeDE considers the uncertainty of pseudotime inference



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 ✉

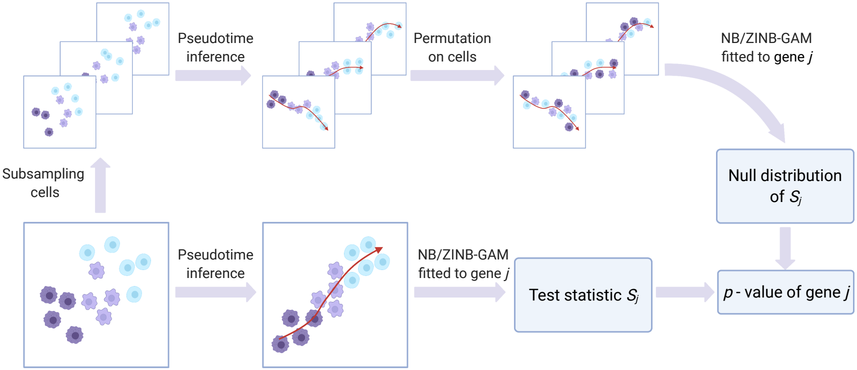
Genome Biology **22**, Article number: 124 (2021) | Cite this article

7705 Accesses | **4** Citations | **29** Altmetric | Metrics

3. Treatment of inferred covariates as observed

Identifying DEGs along inferred pseudotime from single-cell RNA-seq data

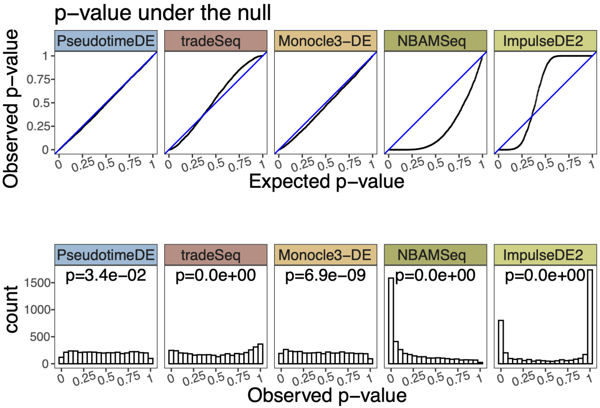
- PseudotimeDE generates well-calibrated p-values for FDR control & uses a generalized additive model (GAM) to achieve good power

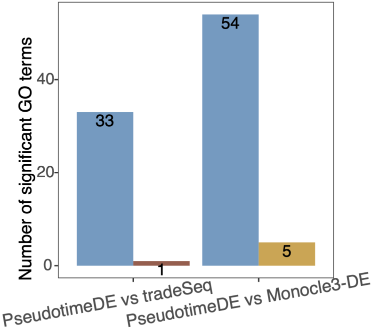


3. Treatment of inferred covariates as observed

Identifying DEGs along inferred pseudotime from single-cell RNA-seq data

- PseudotimeDE generates well-calibrated p-values for FDR control & uses a generalized additive model (GAM) to achieve good power





3. Treatment of inferred covariates as observed

Identifying DEGs along inferred pseudotime from single-cell RNA-seq data

PseudotimeDE limitations

- computational time: high-resolution p-values require $> 10^3$ rounds of (subsampling + pseudotime inference + permutation)

3. Treatment of inferred covariates as observed

Identifying DEGs along inferred pseudotime from single-cell RNA-seq data

PseudotimeDE limitations

- computational time: high-resolution p-values require $> 10^3$ rounds of (subsampling $+$ pseudotime inference $+$ permutation)

  Q: how to reduce the number of rounds while still achieving FDR control?
  A: Clipper

3. Treatment of inferred covariates as observed

Identifying DEGs along inferred pseudotime from single-cell RNA-seq data

PseudotimeDE limitations

- computational time: high-resolution p-values require $> 10^3$ rounds of (subsampling + pseudotime inference + permutation)

  Q: how to reduce the number of rounds while still achieving FDR control?
  A: Clipper

- complete null: what if cells do not follow a trajectory

3. Treatment of inferred covariates as observed

Identifying DEGs along inferred pseudotime from single-cell RNA-seq data

PseudotimeDE limitations

- computational time: high-resolution p-values require $> 10^3$ rounds of (subsampling + pseudotime inference + permutation)

  Q: how to reduce the number of rounds while still achieving FDR control?
  A: Clipper

- complete null: what if cells do not follow a trajectory

  Q: how to generate the null cells?
  A: simulator scDesign3

3. Treatment of inferred covariates as observed

Identifying DEGs along inferred pseudotime from single-cell RNA-seq data

- PseudotimeDE

Identifying DEGs between inferred cell clusters from single-cell RNA-seq data

- ClusterDE (cell clustering + DEG identification between cell clusters)
  - existing methods assume Gaussian distributions

    `TN test` [Zhang, Kamath, and Tse, *Cell Syst*, 2019]

    `clusterpval` [Gao, Bien, and Witten, *arXiv*, 2020]

3. Treatment of inferred covariates as observed

Identifying DEGs along inferred pseudotime from single-cell RNA-seq data

- PseudotimeDE

Identifying DEGs between inferred cell clusters from single-cell RNA-seq data

- ClusterDE (cell clustering + DEG identification between cell clusters)
  - existing methods assume Gaussian distributions

    `TN test` [Zhang, Kamath, and Tse, *Cell Syst*, 2019]

    `clusterpval` [Gao, Bien, and Witten, *arXiv*, 2020]

Our proposal: Clipper + scDesign3

— inspired by

`gap statistic` [Hastie, Tibshirani, and Walther, *JRSSB*, 2002]

`knockoffs` [Barber and Candès, *Ann Stat*, 2015]

Halftime summary

Three common causes of ill-posed p-values

1. Formulation of a two-sample test as a one-sample test
2. Specification of a parametric model that does not fit data well
3. Treatment of inferred covariates as observed

Halftime summary

Three common causes of ill-posed p-values

1. Formulation of a two-sample test as a one-sample test
2. Specification of a parametric model that does not fit data well
3. Treatment of inferred covariates as observed

Clipper: p-value-free FDR control for genomics feature screening
— using FDR control procedure from [Barber and Candès, *Ann Stat*, 2015]

Clipper: p-value-free FDR control for genomics feature screening

- **NO requirement of**
  - high-resolution p-values
  - parametric distributions
  - large sample sizes

- **Foundation: knockoffs**
- **Two components**
  - **contrast scores**
  - **cutoff**

Goal: marginal screening for **interesting** features

d features

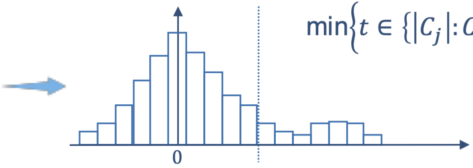
FDR threshold q

Contrast scores

C_1

$\vdots$

C_d



Distribution of $C_1, \ldots, C_d$

Contrast score cutoff

$$\min\left\{t \in \{|C_j| : C_j \neq 0\} : \frac{1+\#\{j: C_j \leq -t\}}{\#\{j: C_j \geq t\} \vee 1} \leq q\right\}$$

Clipper: p-value-free FDR control for genomics feature screening

Key: **contrast score** construction

example	target data	null data
ChIP-seq peak calling (1 vs. 1)	experimental condition	background condition
RNA-seq DEG identification	actual data	permuted data
PseudotimeDE & **ClusterDE**	actual data	**scDesign3** simulated data

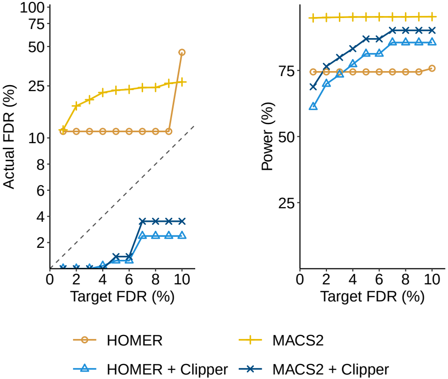
Contrast score of feature $j = 1, \ldots, d$, the

$$C_j := t(\textbf{target data}) - t(\textbf{null data}),$$

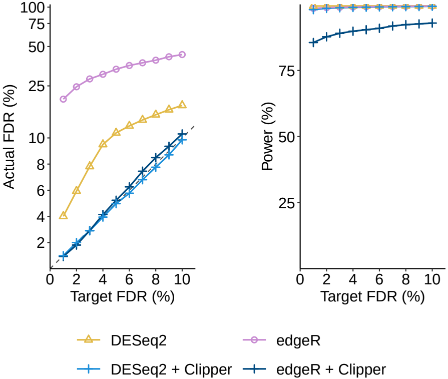
where $t(\cdot)$ is a summary statistic — can be a **complex pipeline**

Clipper rectifies FDR control

ChIP-seq peaking calling

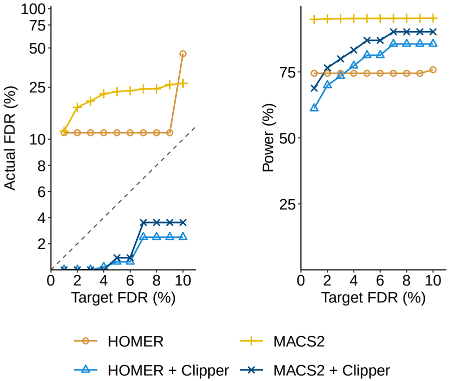


RNA-seq DEG identification

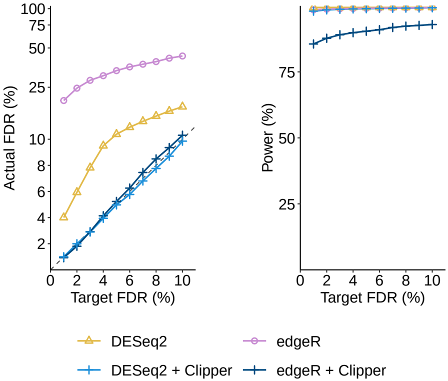


Clipper rectifies FDR control

ChIP-seq peaking calling



RNA-seq DEG identification



Q: how to generate null data to construct contrast scores for PseudotimeDE and ClusterDE?

The last piece

Three common causes of ill-posed p-values

1. Formulation of a two-sample test as a one-sample test

2. Specification of a parametric model that does not fit data well

3. Treatment of inferred covariates as observed

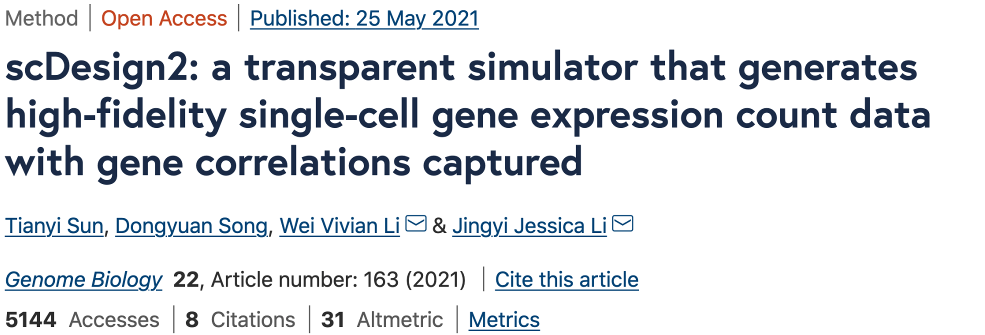
Clipper: a p-value-free FDR control framework

scDesign3: an omnibus single-cell omics simulator

scDesign2: a probabilistic single-cell gene expression data simulator

A multi-gene probabilistic model per cell type

- Each gene $\sim$ count distribution $\in \{\text{Poisson}, \text{negative binomial}, \text{ZIP}, \text{ZINB}\}$
- Gene correlations estimated via Gaussian copula



Method | Open Access | Published: 25 May 2021

scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured

Tianyi Sun, Dongyuan Song, Wei Vivian Li ✉ & Jingyi Jessica Li ✉

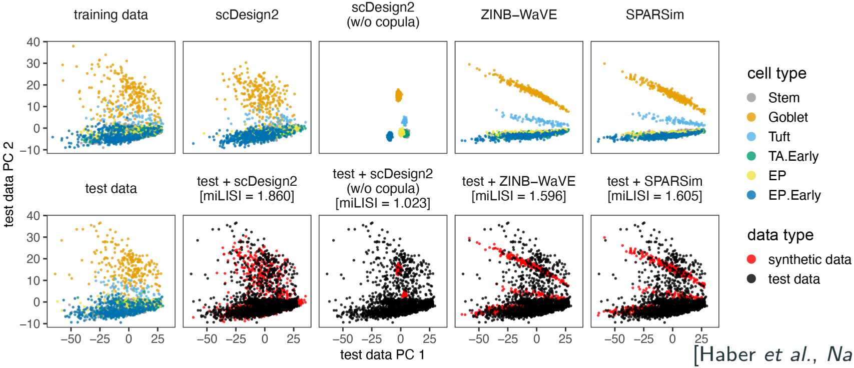
Genome Biology **22**, Article number: 163 (2021) | Cite this article

5144 Accesses | **8** Citations | **31** Altmetric | Metrics

scDesign2: a probabilistic single-cell gene expression data simulator

A multi-gene probabilistic model per cell type

- Each gene $\sim$ count distribution $\in \{\text{Poisson}, \text{negative binomial}, \text{ZIP}, \text{ZINB}\}$
- Gene correlations estimated via Gaussian copula

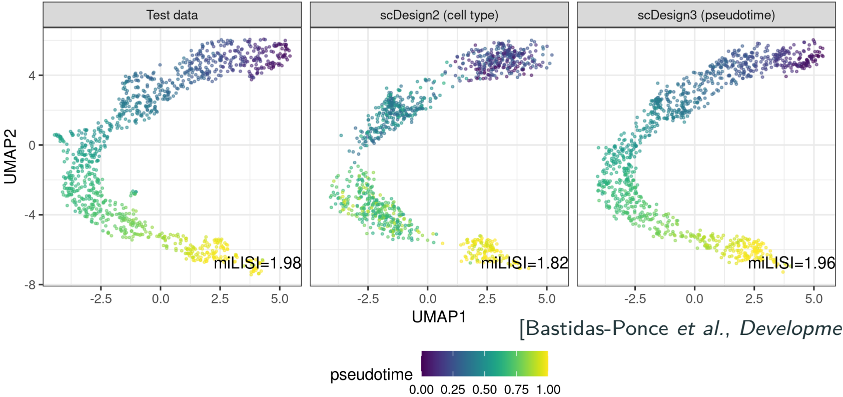


[Haber *et al.*, *Nature*, 2017]

scDesign3: an omnibus single-cell & spatial omics simulator

- **Cell states**: continuous trajectory & discrete cell types
- **Feature modalities**: RNA, ATAC, protein, spatial coordinates, etc.
- **Model selection by likelihood**: vine copula [Joe and Kurowicka's book, 2011]

Example: continuous trajectory (pancreatic cell differentiation)

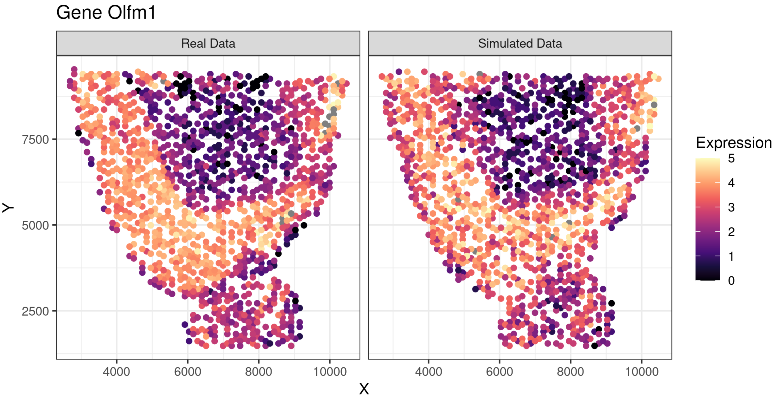


[Bastidas-Ponce *et al.*, *Development*, 2019]

scDesign3: an omnibus single-cell & spatial omics simulator

- **Cell states**: continuous trajectory & discrete cell types
- **Feature modalities**: RNA, ATAC, protein, spatial coordinates, etc.
- **Model selection by likelihood**: vine copula [Joe and Kurowicka's book, 2011]

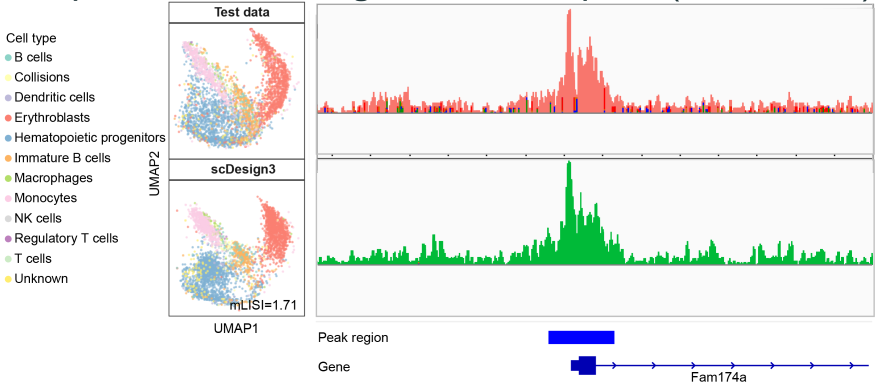
Example: spatial data (brain region measured by 10X Visium)



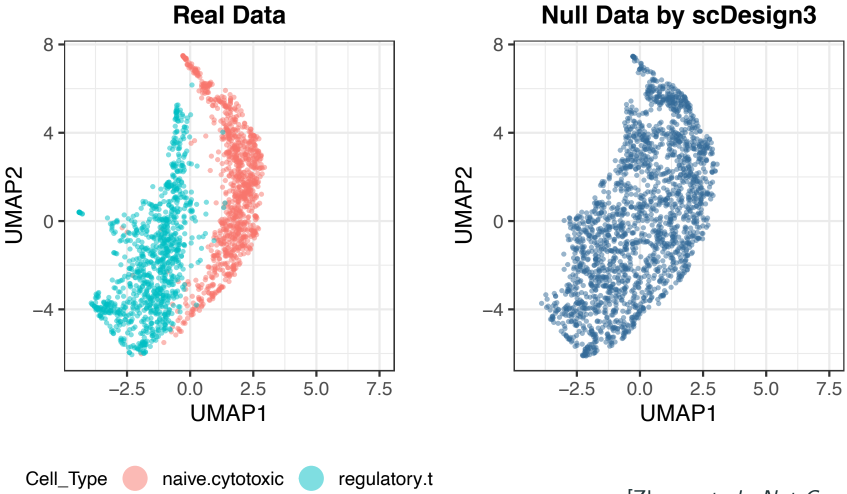
scDesign3: an omnibus single-cell & spatial omics simulator

- **Cell states**: continuous trajectory & discrete cell types
- **Feature modalities**: RNA, ATAC, protein, spatial coordinates, etc.
- **Model selection by likelihood**: vine copula [Joe and Kurowicka's book, 2011]

Example: bone marrow single-cell ATAC-seq data (+ scReadSim)



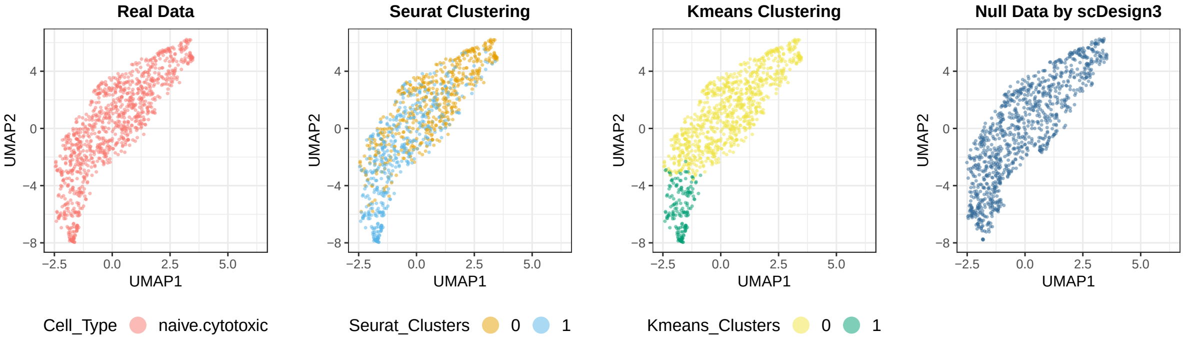
ClusterDE: scDesign3 for null data generation (preliminary)



[Zheng *et al.*, *Nat Commun*, 2017]

ClusterDE: Clipper + scDesign3 (preliminary)

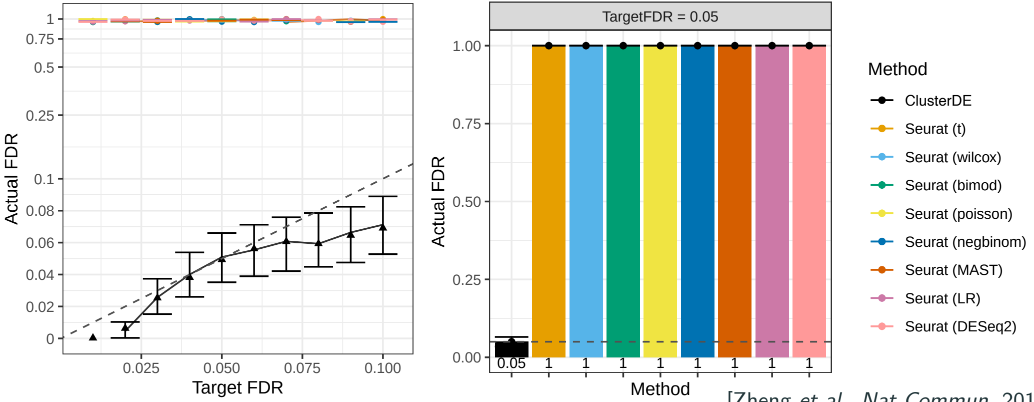
Complete null case: no cell clusters



[Zheng *et al.*, *Nat Commun*, 2017]

ClusterDE: Clipper + scDesign3 (preliminary)

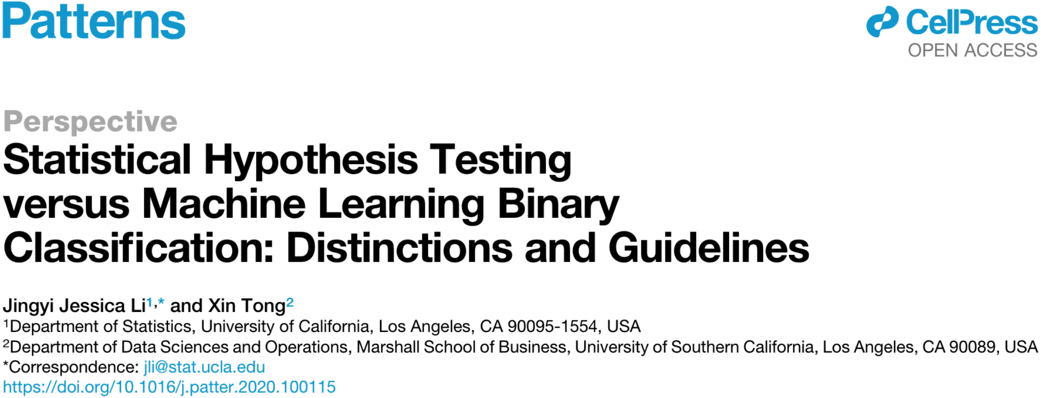
Complete null case: no cell clusters



[Zheng *et al.*, *Nat Commun*, 2017]

Summary: p-values are easily ill-posed in genomics data analysis

Q: should a scientific question be formulated as multiple testing?



Patterns

CellPress
OPEN ACCESS

Perspective

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

Jingyi Jessica Li[1,*] and Xin Tong[2]
[1]Department of Statistics, University of California, Los Angeles, CA 90095-1554, USA
[2]Department 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

Summary: p-values are easily ill-posed in genomics data analysis

Q: should a scientific question be formulated as multiple testing?

If YES, three common causes of ill-posed p-values

1. Formulation of a two-sample test as a one-sample test
   – ChIP-seq peak calling
2. Specification of a parametric model that does not fit data well
   – RNA-seq DEG identification
3. Treatment of inferred covariates as observed
   – single-cell RNA-seq PseudotimeDE & ClusterDE

Summary: p-values are easily ill-posed in genomics data analysis

Q: should a scientific question be formulated as multiple testing?

If YES, three common causes of ill-posed p-values

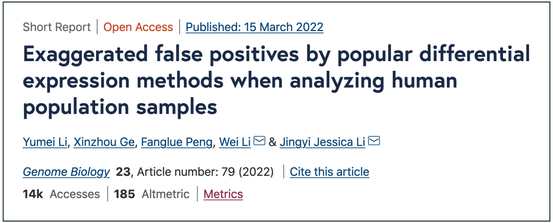
1. Formulation of a two-sample test as a one-sample test
   – ChIP-seq peak calling
2. Specification of a parametric model that does not fit data well
   – RNA-seq DEG identification
3. Treatment of inferred covariates as observed
   – single-cell RNA-seq PseudotimeDE & ClusterDE

Clipper: a p-value-free FDR control framework

scDesign3: an omnibus single-cell & spatial omics simulator

– fair benchmarking of computational tools ($>$ 1000 at `www.scrna-tools.org`)

Summary: relevant publications



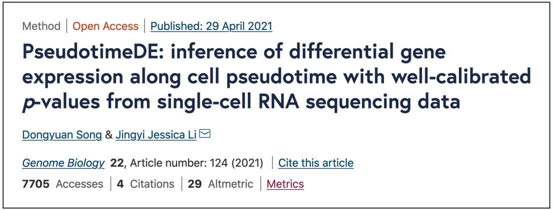
Short Report | Open Access | Published: 15 March 2022

Exaggerated false positives by popular differential expression methods when analyzing human population samples

Yumei Li, Xinzhou Ge, Fanglue Peng, Wei Li ✉ & Jingyi Jessica Li ✉

Genome Biology **23**, Article number: 79 (2022) | Cite this article

14k Accesses | **185** Altmetric | Metrics



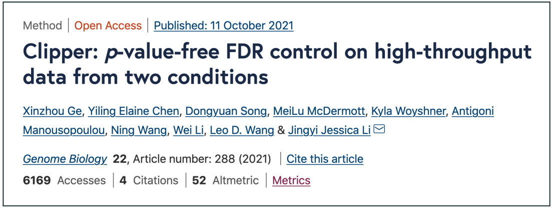
Method | Open Access | Published: 29 April 2021

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

7705 Accesses | **4** Citations | **29** Altmetric | Metrics



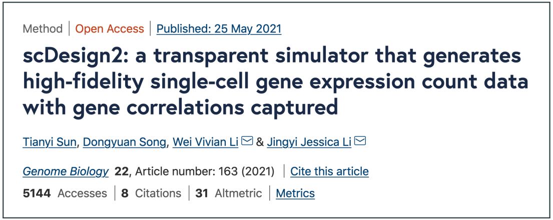
Method | Open Access | Published: 11 October 2021

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

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

Genome Biology **22**, Article number: 288 (2021) | Cite this article

6169 Accesses | **4** Citations | **52** Altmetric | Metrics



Method | Open Access | Published: 25 May 2021

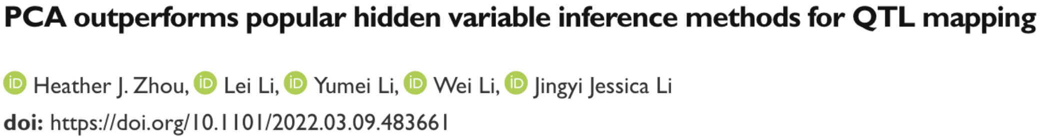
scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured

Tianyi Sun, Dongyuan Song, Wei Vivian Li ✉ & Jingyi Jessica Li ✉

Genome Biology **22**, Article number: 163 (2021) | Cite this article

5144 Accesses | **8** Citations | **31** Altmetric | Metrics

A reviewer’s quote



PCA outperforms popular hidden variable inference methods for QTL mapping

Heather J. Zhou, Lei Li, Yumei Li, Wei Li, Jingyi Jessica Li

doi: https://doi.org/10.1101/2022.03.09.483661

“These results may come as a surprise to some, given the nearly uncontestable status that *method A* has achieved within the community, but sadly they reflect the fact that computational biology methods can rise to fame almost **by accident rather than by sound statistical arguments**.”

Acknowledgements

Ph.D. advisors @ Berkeley: **Peter J. Bickel** & **Haiyan Huang**

Acknowledgements

Ph.D. advisors @ Berkeley: **Peter J. Bickel** & **Haiyan Huang**

Dr. Yumei Li
(Collaborator postdoc @ UCI)
DE genes

Dr. Wei Li
(Collaborator PI @ UCI)
Clipper
DE genes

Dr. Xinzhou Ge
(Postdoc)
Clipper
DE genes

Dr. Yiling Elaine Chen
(Former Ph.D. student)
Clipper

Tianyi Sun
(Ph.D. student)
scDesign2

Dongyuan Song
(Ph.D. student)
PseudotimeDE
scDesign3

NIH NIGMS
NSF
Alfred P. Sloan Foundation
Johnson & Johnson WiSTEM²D Women in Science, Technology, Engineering, Math, Manufacturing and Design
W.M. Keck Foundation
PhRMA Research • Progress • Hope