UCLA Statistics

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

Jingyi Jessica Li

Joint work with Xinzhou Ge and Yiling Elaine Chen (Ph.D. students)

Department of Statistics
University of California, Los Angeles

http://jsb.ucla.edu

Introduction

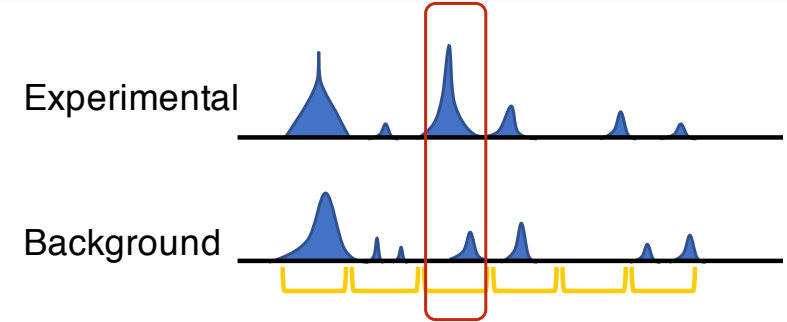
Background

- High-throughput biological data
  - Small sample size (number of replicates, often ≤ 3)
  - Huge number of features (often $\sim 10^4$)
- Two conditions
  - Experimental
  - Background / negative control
- Identification of "interesting" features

Four example high-throughput biological applications

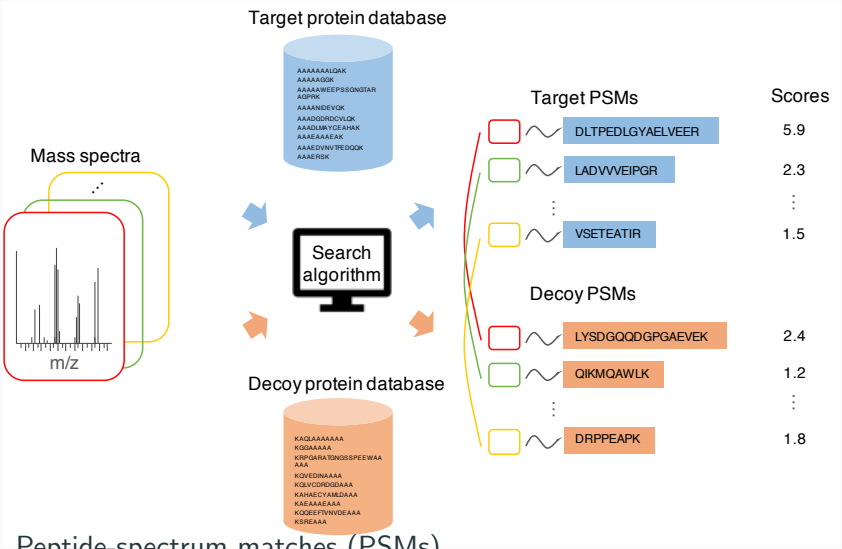
- Peak calling from ChIP-seq data
  - Protein-DNA binding sites
- Peptide identification from mass spectrometry (MS) data
  - Peptide-spectrum matches
- Differential analysis of RNA-seq data
  - Differentially expressed genes
- Differential analysis of Hi-C data
  - Differentially interacting chromatin regions

Peak calling from ChIP-seq data



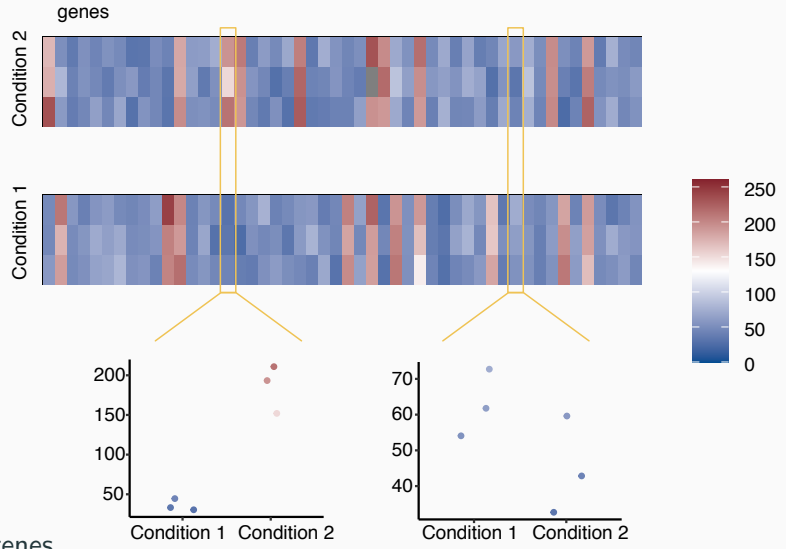
Features: genomic regions

Peptide identification from MS data

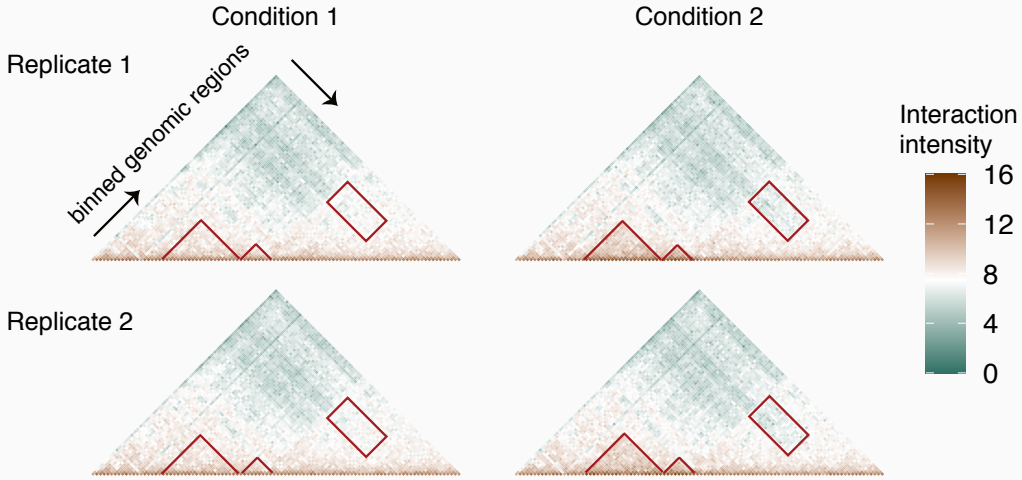


Features: Peptide-spectrum matches (PSMs)

Identification of differentially expressed genes from RNA-seq data



Identification of differentially interacting regions from Hi-C data



Features: region pairs

Enrichment and differential analyses

- Interesting means "**enriched**" or "**differential**"
- Enriched features:

$$\mathbb{E}[\text{experimental}] > \mathbb{E}[\text{background}]$$

  - Protein binding regions (peaks)
  - Peptide-spectrum matches

- Differential features:

$$\mathbb{E}[\text{experimental}] \neq \mathbb{E}[\text{background}]$$

  - Differentially expressed genes
  - Differentially interacting regions

False discovery rate (FDR)

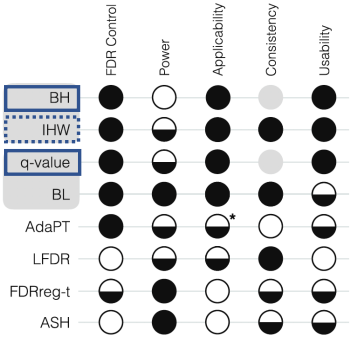
- Criterion for controlling false discoveries
- **Frequentist FDR** [Benjamini and Hochberg, *JRSSB*, 1995]

$$\text{FDR} := \mathbb{E}\left[\frac{\#\text{ false discoveries}}{\#\text{ discoveries } \vee 1}\right]$$

- Bayesian paradigm:
  - Bayesian false discovery rate [Efron and Tibshirani, *Genet Epidemiol*, 2002]
  - Local false discovery rate (**fdr**) [Efron et al., *JASA*, 2001]
  - Local false sign rate [Stephens, *Biostatistics*, 2017]

Existing FDR control methods

	Input	Assumptions	Output	R package
BH	*p*-values	exchangeability	adjusted *p*-values	`stats`
IHW	(1) *p*-values (2) independent & informative covariate	exchangeability within covariate groups		`ihw`
q-value	*p*-values	exchangeability	*q*-values	`qvalue`
BL	(1) *p*-values (2) independent & informative covariate	exchangeability conditional on covariate(s)	adjusted *p*-values	`swfdr`
AdaPT			*q*-values	`adaptMT`
LFDR		exchangeability within covariate groups	adjusted *p*-values	none
FDRreg	(1) *z*-scores (2) independent & informative covariate	exchangeability conditional on covariate(s); normal test statistics	Bayesian FDRs	`FDRreg`
ASH	(1) effect sizes (2) standard errors of (1)	effects are unimodal; test statistics have normal or *t* mixture components	*q*-values	`ash`



"A practical guide to methods controlling false discoveries in computational biology" [Korthauer et al., *Genome Biol*, 2019]

Generic FDR control methods

- **P-value based methods** (exact)
  - Benjamini-Hochberg (**BH**) procedure [Benjamini and Hochberg, *JRSSB*, 1995]
  - Storey's **q-value** procedure [Storey, *JRSSB*, 2002]
- **Local fdr based method** (approximate)
  - Thresholding local fdr to q (e.g., 5%) approximately controls the FDR

Calculation of p-values

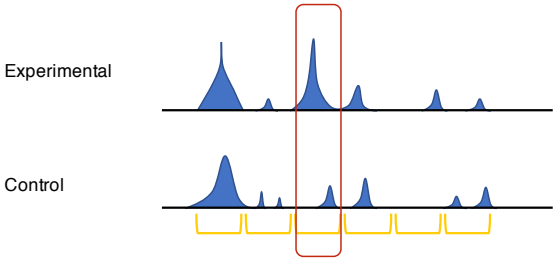
- Requirements
  - Distributional assumptions (parametric)
  - Large number of replicates (nonparametric)
- Approaches
  - "**Paired**" approach
  - "**Pooled**" approach

Paired approach

- Used in
  - Peak calling from ChIP-seq data
  - Identification of differentially expressed genes from RNA-seq data
  - Identification of differentially interacting regions from Hi-C data
- One feature at a time, two-sample test

- Issues
  - Mis-formulation (e.g., two-sample test as one-sample test)
  - Mis-specification (e.g., negative binomial as Poisson)

Example of paired mis-formulation: MACS [Zhang et al., Genome Biol, 2008]



In the control samples, we often observe tag distributions with local fluctuations and biases. For example, at the FoxA1 candidate peak locations, tag counts are well correlated between ChIP and control samples (Figure 1c,d). Many possible sources for these biases include local chromatin structure, DNA amplification and sequencing bias, and genome copy number variation. Therefore, instead of using a uniform λ_{BG} estimated from the whole genome, MACS uses a dynamic parameter, λ_{local}, defined for each candidate peak as:

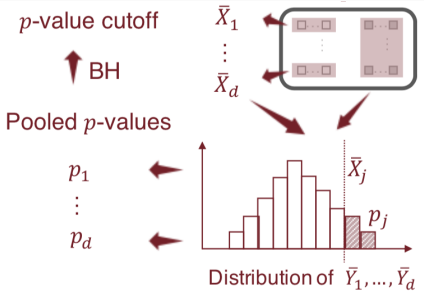
$$\lambda_{local} = \max(\lambda_{BG}, [\lambda_{1k}, \lambda_{5k}, \lambda_{10k})$$

where λ_{1k}, λ_{5k} and λ_{10k} are λ estimated from the 1 kb, 5 kb or 10 kb window centered at the peak location in the control sample, or the ChIP-Seq sample when a control sample is not available (in which case λ_{1k} is not used). λ_{local} captures the influence of local biases, and is robust against occasional low tag counts at small local regions. MACS uses λ_{local} to calculate the *p*-value of each candidate peak and removes potential false positives due to local biases (that is, peaks significantly under λ_{BG}, but not under λ_{local}). Candidate peaks with *p*-values below a user-defined threshold *p*-value (default 10^{-5}) are called, and the ratio between the ChIP-Seq tag count and λ_{local} is reported as the *fold_enrichment*.

Cited for more than 8,000 times

Pooled approach

- Used in
  - Peptide identification from MS data
- Pools all features' background measurements to form a null distribution
  - Assumes a **homogeneous** background: features are i.i.d. under background



Generic FDR control methods

- **P-value based methods** (exact)
  - Benjamini-Hochberg (**BH**) procedure [Benjamini and Hochberg, *JRSSB*, 1995]
  - Storey's **q-value** procedure [Storey, *JRSSB*, 2002]
- **Local fdr based method** (approximate)
  - Thresholding local fdr to q (e.g., 5%) approximately controls the FDR

Local fdr [Efron et al., JASA, 2001]

- Local fdr of feature j

$$\text{local fdr}_j := \mathbb{P}(\text{feature is uninteresting} \mid Z = z_j)$$

  vs. Bayesian false discovery rate

$$\text{Fdr}(z) := \mathbb{P}(\text{feature is uninteresting} \mid Z \geq z)$$

- It can be shown that (assuming local fdr_j is monotone decreasing in z_j)

$$\text{Fdr}(z^*) \leq q \text{ if } z^* := \min\{z_j : \text{local fdr}_j \leq q\}$$

- With discoveries $:= \{\text{feature } j : \text{local fdr}_j \leq q\}$

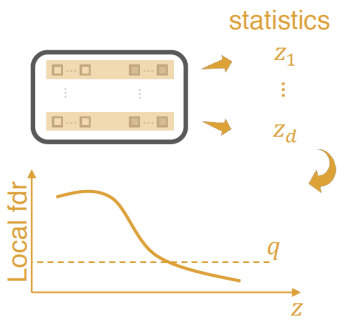
$$\text{FDR}(\text{discoveries}) \approx \text{Fdr}(z^*)$$

- So FDR is approximately controlled

Local fdr [Efron et al., JASA, 2001]

- Requires estimating the null distribution of test statistic by
  - Normal distributional assumption

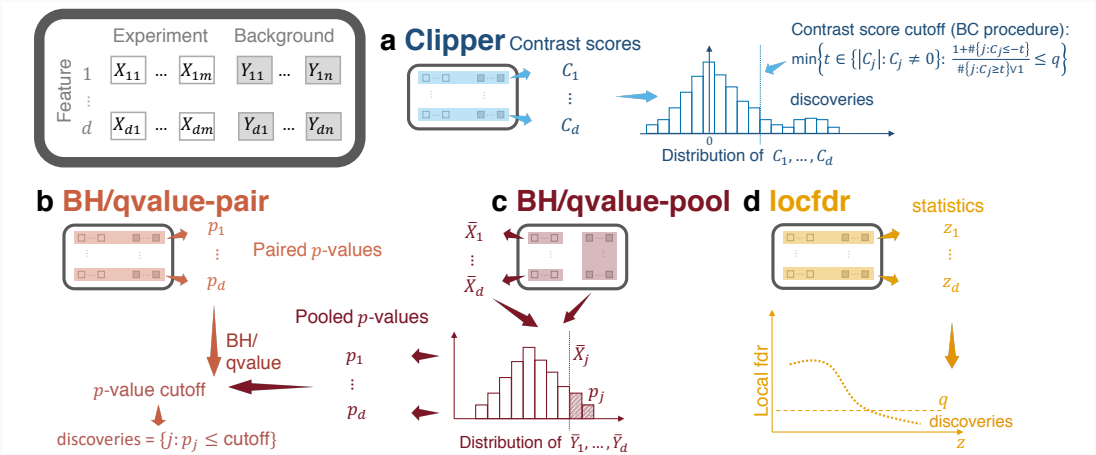
    or
  - Swapping replicates between conditions



Our proposal: Clipper

- **Does not**
  - use p-values
  - assume parametric distributions
  - require many replicates
- Two components
  - **Contrast scores**
  - **Cutoff**
    - Enrichment analysis with equal numbers of replicates: **BC procedure** [Barber and Candès, *Ann Stat*, 2015]
    - Differential analysis and other enrichment analysis: **GZ procedure** [Gimenez and Zou, *PMLR*, 2019]
- Robust to
  - distributions
  - numbers of replicates
  - outliers

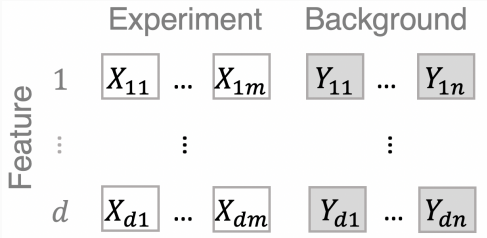
Illustration of methods



Clipper Method

Notations

- d: number of features
- $\boldsymbol{X}_j = (X_{j1}, \ldots, X_{jm})^\top$, $\boldsymbol{Y}_j = (Y_{j1}, \ldots, Y_{jn})^\top$: measurements under two conditions
  - $X_{j1}, \ldots, X_{jm} \geq 0$ are i.i.d.; $\mu_{Xj} = \mathbb{E}[X_{j1}]$
  - $Y_{j1}, \ldots, Y_{jn} \geq 0$ are i.i.d.; $\mu_{Yj} = \mathbb{E}[Y_{j1}]$
  - $j = 1, \ldots, d$



Feature j is **interesting**

- Enrichment analysis: $\mu_{Xj} > \mu_{Yj}$
- Differential analysis: $\mu_{Xj} \neq \mu_{Yj}$

Assumptions

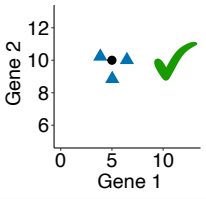
Conditioning on $\{\mu_{Xj}\}_{j=1}^{d}$ and $\{\mu_{Yj}\}_{j=1}^{d}$,

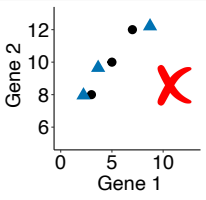
- **Independence**:

$$X_{j1}, \ldots, X_{jm}, Y_{j1}, \ldots, Y_{jn} \text{ are mutually independent} \quad (1)$$
$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_d \text{ are mutually independent}$$

- For **uninteresting** feature j,

$$X_{j1}, \ldots, X_{jm}, Y_{j1}, \ldots, Y_{jn} \text{ are identically distributed} \quad (2)$$





Expectation
Measurement

Clipper for three analysis tasks

Calculation of contrast scores depends on analysis tasks:

- Enrichment analysis with $m = n$
- Enrichment analysis with $m \neq n$
- Differential analysis

Enrichment analysis with $m = n$: contrast scores

Two summary statistics:

$$t^{\text{diff}}(\boldsymbol{x}, \boldsymbol{y}) := \bar{x} - \bar{y} \tag{3}$$

$$t^{\max}(\boldsymbol{x}, \boldsymbol{y}) := \max(\bar{x}, \bar{y}) \cdot \operatorname{sign}(\bar{x} - \bar{y}) \tag{4}$$

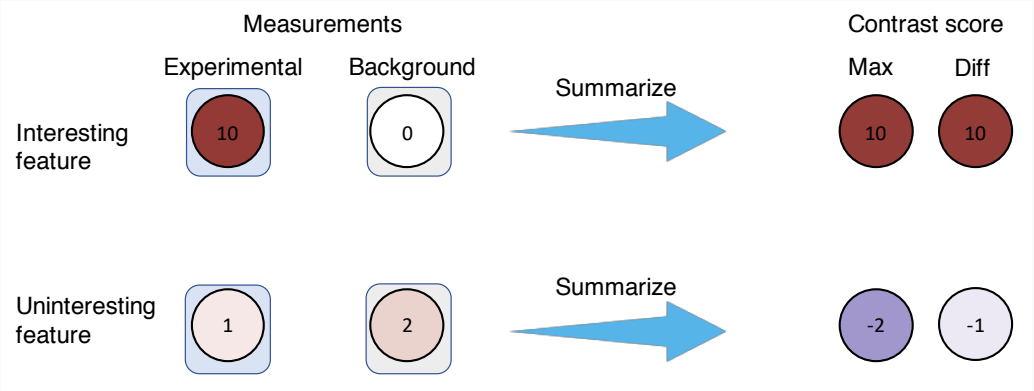
In enrichment analysis with $m = n$, **contrast score** of feature j:

$$C_j := t^{\text{diff}}(\boldsymbol{X}_j, \boldsymbol{Y}_j) \qquad \textbf{difference} \text{ contrast score} \tag{5}$$

or

$$C_j := t^{\max}(\boldsymbol{X}_j, \boldsymbol{Y}_j) \qquad \textbf{maximum} \text{ contrast score} \tag{6}$$

Enrichment analysis with $m = n = 1$: contrast scores

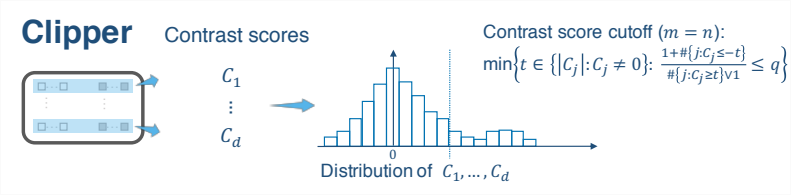


Enrichment analysis with $m = n$: cutoff

Definition 1 BC procedure [Barber and Candès, Ann Stat, 2015]

- Given contrast scores $\{C_j\}_{j=1}^d$, define $\mathcal{C} = \{|C_j| : C_j \neq 0 \, ; \, j = 1, \ldots, d\}$
- Based on the target FDR threshold $q \in (0, 1)$, contrast-score cutoff T^{BC}:

$$T^{\mathrm{BC}} := \min \left\{ t \in \mathcal{C} : \frac{\mathrm{card}(\{j : C_j \leq -t\}) + 1}{\mathrm{card}(\{j : C_j \geq t\}) \vee 1} \leq q \right\} \tag{7}$$



- **Discoveries**: $\{j : C_j \geq T^{\mathrm{BC}}\}$
- BC vs. BH [Arias-Castro and Chen, *Electronic J Stat*, 2016]

Key lemma to guarantee the theoretical FDR control by the BC procedure

Define $S_j = \operatorname{sign}(C_j) \in \{-1, 0, 1\}$

$\mathcal{N}$: the set of uninteresting features

Then

1. $S_1, \ldots, S_d$ are mutually independent
2. $\mathbb{P}(S_j = 1) = \mathbb{P}(S_j = -1)$ for all $j \in \mathcal{N}$
3. $\{S_j\}_{j \in \mathcal{N}} \perp \mathcal{C}$

When $m \neq n$, 2 and 3 are not guaranteed to hold

Enrichment analysis with $m \neq n$: permutation

Data:

$$\mathbf{W} := \begin{bmatrix} X_{11} & \cdots & X_{1m} & Y_{11} & \cdots & Y_{1n} \\ & \vdots & & & \vdots & \\ X_{d1} & \cdots & X_{dm} & Y_{d1} & \cdots & Y_{dn} \end{bmatrix}$$

- σ: a permutation function to permute the columns of $\mathbf{W}$ and output $\mathbf{W}^\sigma$

$$\mathbf{W}^\sigma := \begin{bmatrix} X^\sigma_{11} & \cdots & X^\sigma_{1m} & Y^\sigma_{11} & \cdots & Y^\sigma_{1n} \\ & \vdots & & & \vdots & \\ X^\sigma_{d1} & \cdots & X^\sigma_{dm} & Y^\sigma_{d1} & \cdots & Y^\sigma_{dn} \end{bmatrix}$$

- The permuted measurements $\left\{ (\boldsymbol{X}^\sigma_j, \boldsymbol{Y}^\sigma_j) \right\}_{j=1}^{d}$

Enrichment analysis with $m \neq n$: permutation

- The identity permutation σ_0
- Sample h non-identity permutations $\sigma_1, \ldots, \sigma_h$
- $\left\{(\boldsymbol{X}_j^{\sigma_0}, \boldsymbol{Y}_j^{\sigma_0}), (\boldsymbol{X}_j^{\sigma_1}, \boldsymbol{Y}_j^{\sigma_1}), \cdots, (\boldsymbol{X}_j^{\sigma_h}, \boldsymbol{Y}_j^{\sigma_h})\right\}_{j=1}^{d}$
- Degree of "**interestingness**" of feature j given σ_ℓ: $T_j^{\sigma_\ell} := t^{\text{diff}}(\boldsymbol{X}_j^{\sigma_\ell}, \boldsymbol{Y}_j^{\sigma_\ell})$
- Sort $\{T_j^{\sigma_\ell}\}_{\ell=0}^{h}$ so that

$$T_j^{(0)} \geq T_j^{(1)} \geq \cdots \geq T_j^{(h)}$$

Enrichment analysis with $m \neq n$: contrast scores

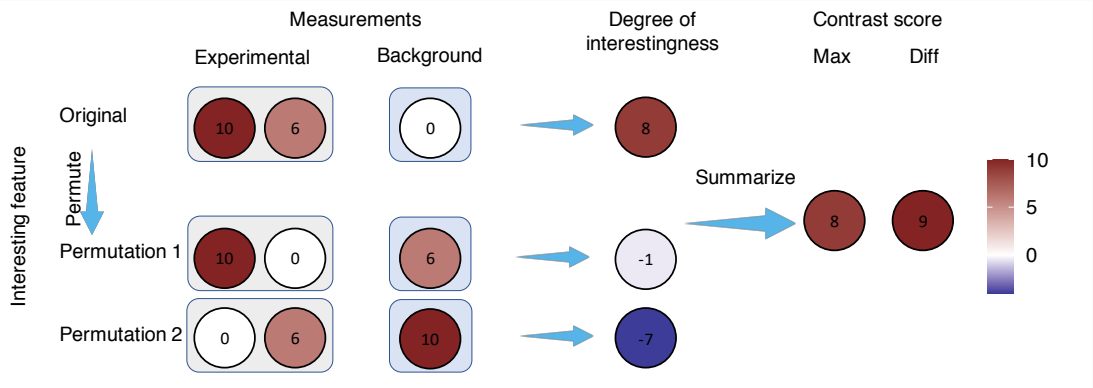
Contrast score of feature j:

$$C_j := \begin{cases} T_j^{(0)} - T_j^{(1)} & \text{if } T_j^{(0)} = T_j^{\sigma_0} \\ T_j^{(1)} - T_j^{(0)} & \text{otherwise} \end{cases} \qquad \textbf{difference} \text{ contrast score} \qquad (8)$$

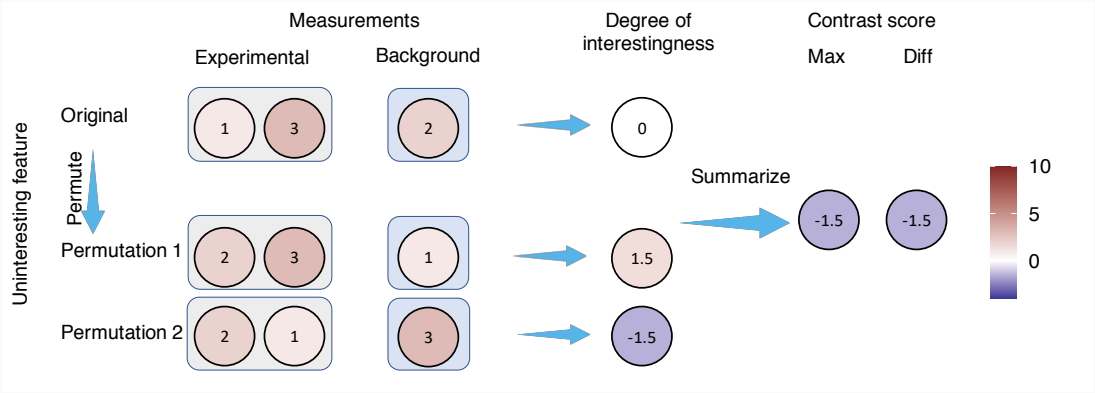
or

$$C_j := \begin{cases} \left| T_j^{(0)} \right| & \text{if } T_j^{(0)} = T_j^{\sigma_0} > T_j^{(1)} \\ 0 & \text{if } T_j^{(0)} = T_j^{(1)} \\ -\left| T_j^{(0)} \right| & \text{otherwise} \end{cases} \qquad \textbf{maximum} \text{ contrast score} \qquad (9)$$

Enrichment analysis with $m = 2$, $n = 1$: contrast scores



Enrichment analysis with $m = 2$, $n = 1$: contrast scores



Definition 2 GZ procedure [Gimenez and Zou, PMLR, 2019]

- Given contrast scores $\{C_j\}_{j=1}^d$, define $\mathcal{C} = \{|C_j| : C_j \neq 0 \, ; \, j = 1, \ldots, d\}$
- Based on the target FDR threshold $q \in (0, 1)$, contrast-score cutoff T^{GZ}:

$$T^{\mathrm{GZ}} := \min \left\{ t \in \mathcal{C} : \frac{\frac{1}{h} + \frac{1}{h}\mathrm{card}\left(\{j : C_j \leq -t\}\right)}{\mathrm{card}\left(\{j : C_j \geq t\}\right) \vee 1} \leq q \right\} \tag{10}$$

- **Discoveries**: $\left\{j : C_j \geq T^{\mathrm{GZ}}\right\}$

Key lemma to guarantee the theoretical FDR control by the GZ procedure

Define $S_j = \mathrm{sign}\,(C_j) \in \{-1, 0, 1\}$

$\mathcal{N}$: the set of uninteresting features

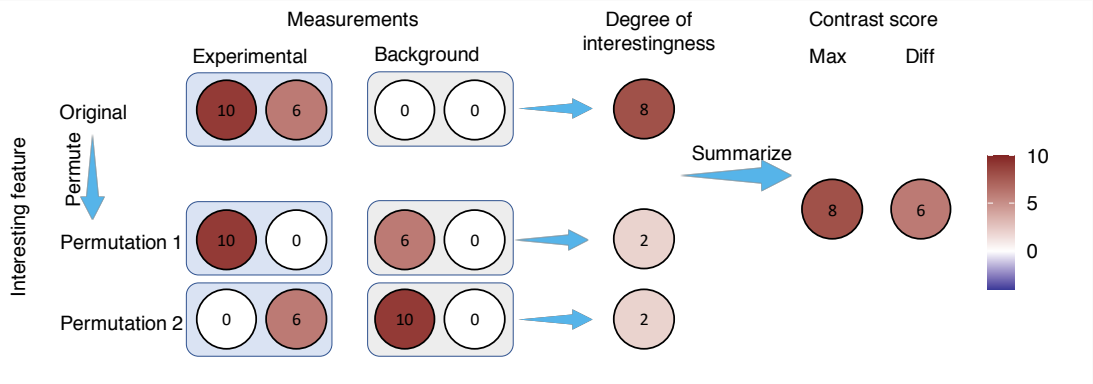
Then

1. $S_1, \ldots, S_d$ are mutually independent ;
2. $\mathbb{P}(S_j = 1) \leq \frac{1}{h+1}$ for all $j \in \mathcal{N}$;
3. $\{S_j\}_{j \in \mathcal{N}} \perp \mathcal{C}$.

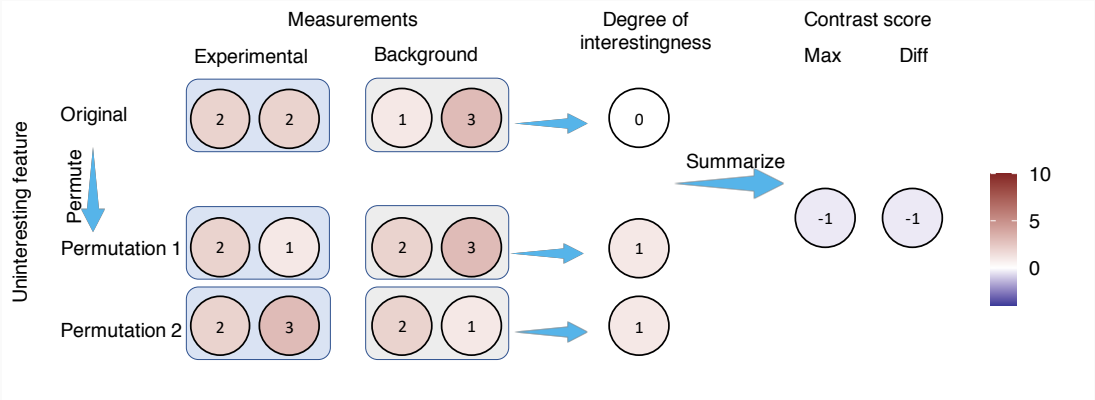
Differential analysis

- Almost the same as enrichment analysis with $m \neq n$
- Only difference: the degree of "interestingness" of feature j:
  - **Differential**:
    $$T_j^{\sigma_\ell} := \left| t^{\text{diff}}(\boldsymbol{X}_j^{\sigma_\ell}, \boldsymbol{Y}_j^{\sigma_\ell}) \right|$$
  - Enrichment:
    $$T_j^{\sigma_\ell} := t^{\text{diff}}(\boldsymbol{X}_j^{\sigma_\ell}, \boldsymbol{Y}_j^{\sigma_\ell})$$
- Contrast-score cutoff also by the GZ procedure

Differential analysis with $m = n = 2$: contrast scores



Differential analysis with $m = n = 2$: contrast scores



Simulation analysis

Simulation design

- **Analysis task**: enrichment or differential
- **Numbers of replicates** *m*vs*n*: 1vs1, 2vs1, 3vs3, or 10vs10
- **Distribution**: Gaussian, Poisson, or negative binomial
- **Background**: homogeneous or heterogeneous
- **Number of features**: $d = 1{,}000$ or $10{,}000$

Generic FDR control methods

- **P-value based methods**
  - p-value calculation approach (paired or pooled)
  - p-value thresholding procedure (BH or Storey's qvalue)

$\Rightarrow$ BH-pair, BH-pool, qvalue-pair, qvalue-pool

- **Local fdr based methods**
  - empirical null (Gaussian)
  - swapping

$\Rightarrow$ locfdr-emp, locfdr-swap

Simulation results

1vs1 Enrichment

Gaussian

d = 1000

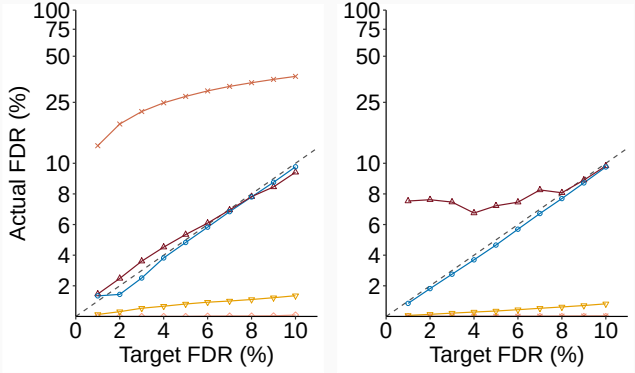
Homogeneous

1vs1 Enrichment

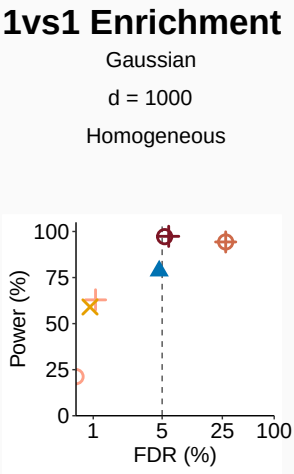
Gaussian

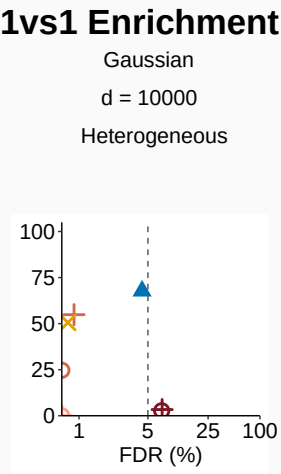
d = 10000

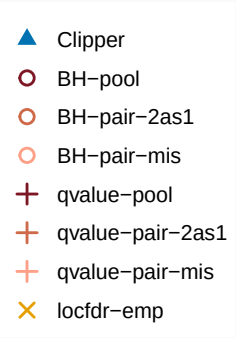
Heterogeneous



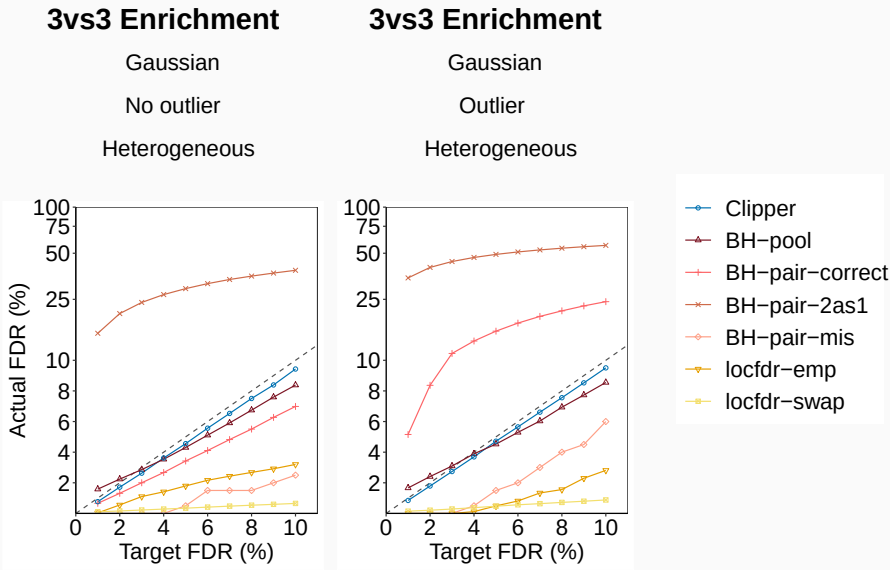
Simulation results: target FDR threshold $q = 5\%$



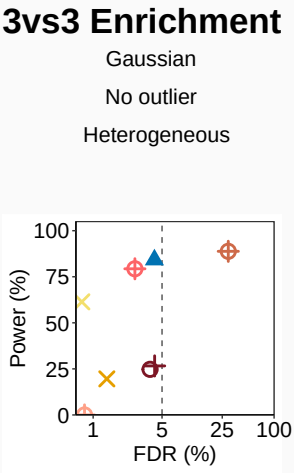


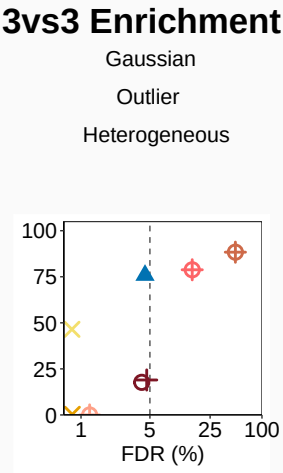


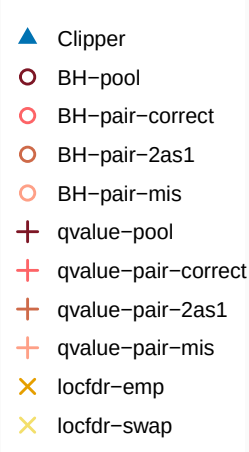
Simulation results



Simulation results: target FDR threshold $q = 5\%$







Simulation results

2vs1 Enrichment

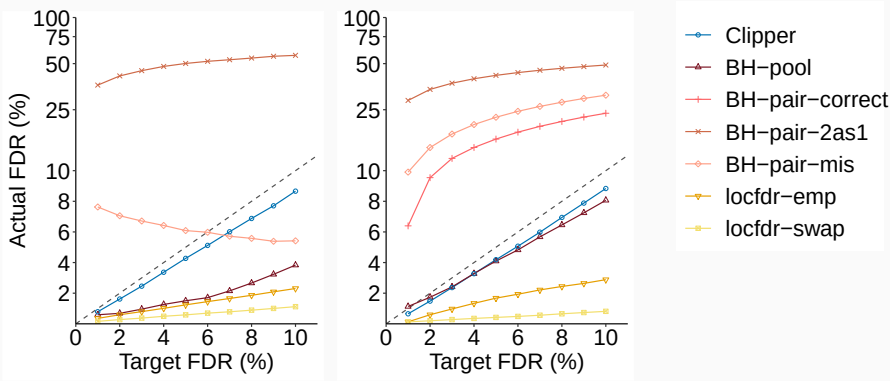
Poisson

Heterogeneous

3vs3 Differential

Negative binomial

Heterogeneous

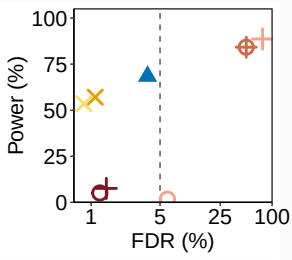


Simulation results: target FDR threshold $q = 5\%$

2vs1 Enrichment

Poisson

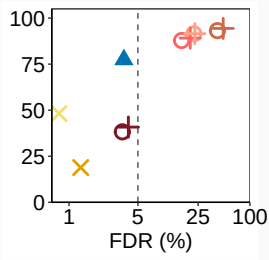
Heterogeneous

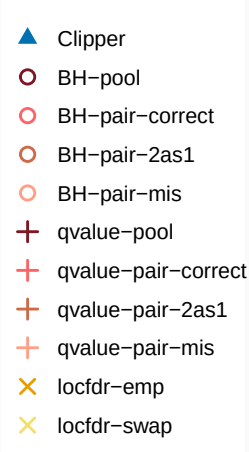


3vs3 Differential

Negative binomial

Heterogeneous





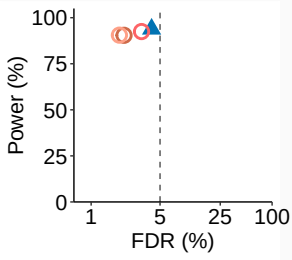
Simulation results: target FDR threshold $q = 5\%$

10vs10 Enrichment

Poisson

No outlier

Heterogeneous

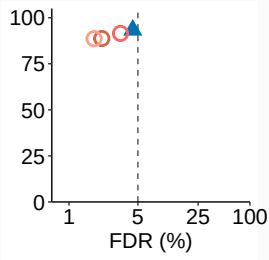


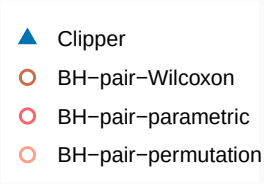
10vs10 Enrichment

Negative Binomial

No Outlier

Heterogeneous





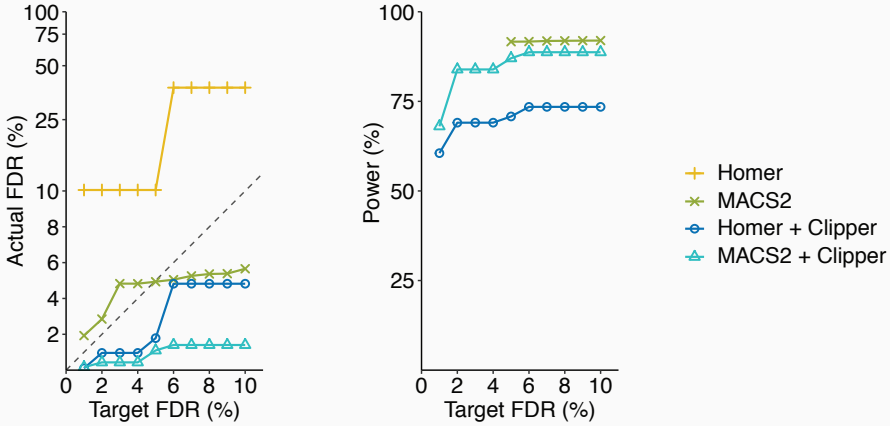
High-throughput biological data applications

High-throughput biological data applications

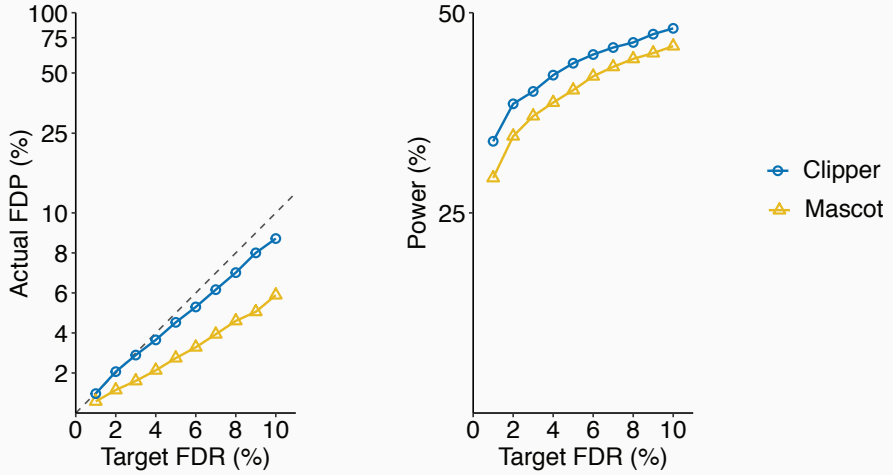
We compare Clipper with mainstream bioinformatics methods

- Peaking calling from ChIP-seq data:
  - **MACS2** [Zhang et al., *Genome Biol*, 2008]
  - **Homer** [Heinz et al., *Mol Cell*, 2010]
- Peptide identification from MS data
  - **Mascot** [Spivak et al.,*J Proteome Res*, 2009]
- Identification of differentially expressed genes from RNA-seq data
  - **DESeq2** [Love et al., *Genome Biol*, 2014]
  - **edgeR** [Robinson et al., *Bioinformatics*, 2010]
  - Covariate-based p-value weighting: **IHW** [Ignatiadis et al., *Nat Methods*, 2016]
- Identification of differentially interacting regions from Hi-C data
  - **diffHic** [Lun et al., *BMC Bioinformatics*, 2015]
  - **FIND** [Djekidel et al., *Genome Res*, 2018]
  - **multiHiCcompare** [Stansfield et al., *Bioinformatics*, 2019]

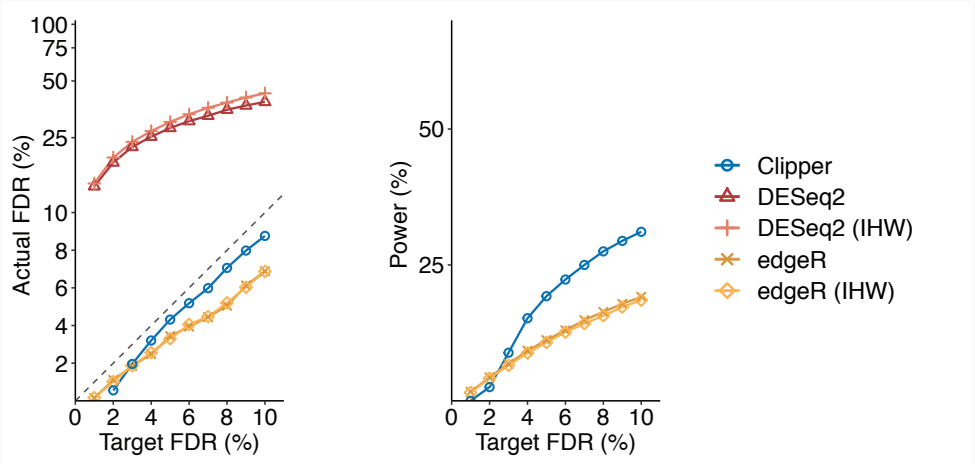
Peaking calling from ChIP-seq data



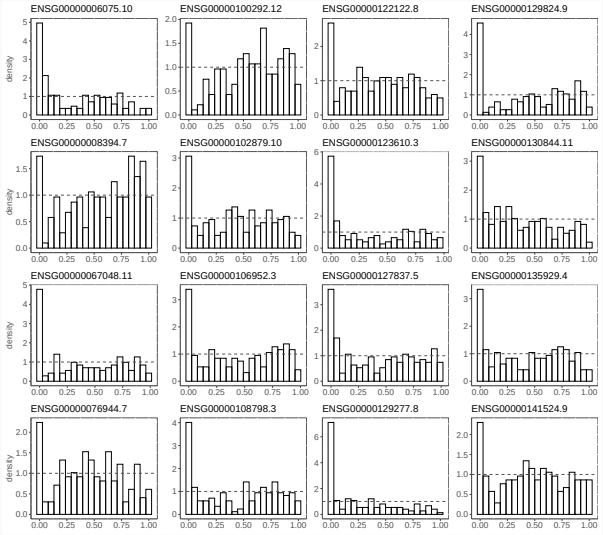
Peptide identification from MS data



Identification of differentially expressed genes from RNA-seq data

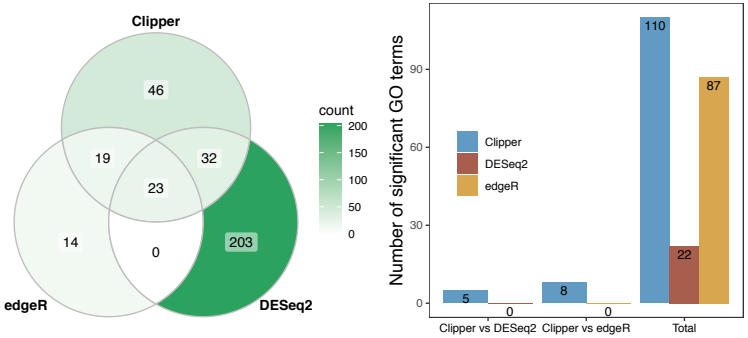


The p-value distributions of 16 non-DEGs most frequently identified by DESeq2



Human monocytes RNA-seq [Williams et al., BMC Bioinformatics, 2017]

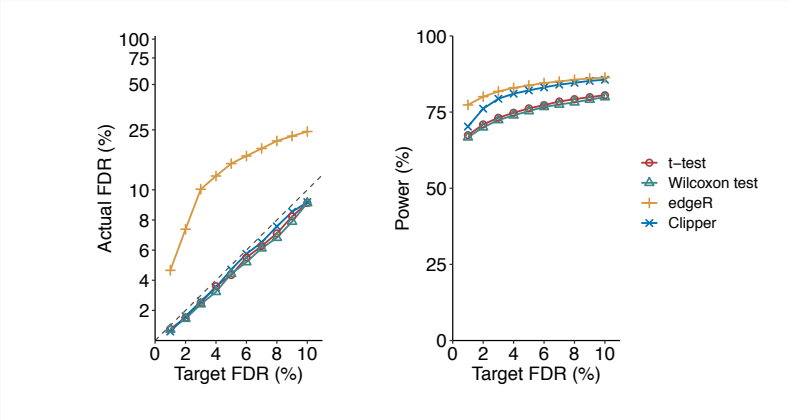
Classical (inflammatory) vs. non-classical (chronical) monocytes



Most significant GO term from DESeq2

GO term (ID)	qvalue (DESeq2)	qvalue (edgeR)	qvalue (Clipper)
leukocyte chemotaxis (GO:0030595)	9.930044e−06	9.594885e−09	3.104557e−10
myeloid leukocyte migration (GO:0097529)	1.107612e−05	2.921486e−08	5.740217e−10
granulocyte chemotaxis (GO:0071621)	2.698853e−05	1.008808e−08	1.167108e−09
neutrophil chemotaxis (GO:0030593)	2.698853e−05	2.921486e−08	2.691033e−09

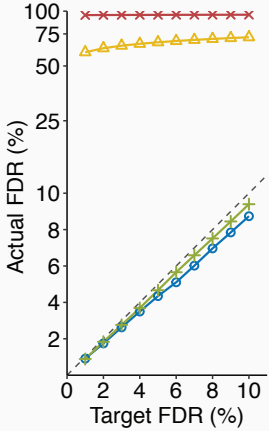
Identification of differentially expressed genes from single-cell RNA-seq data

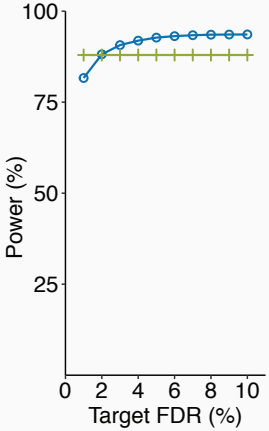


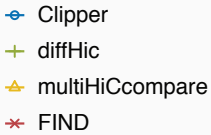
10x Genomics data generated by **scDesign2** https://github.com/JSB-UCLA/scDesign2

"Bias, robustness and scalability in single-cell differential expression analysis" [Soneson and Robinson *Nat Methods*, 2018]

Identification of differentially interacting regions from Hi-C data







Discussion

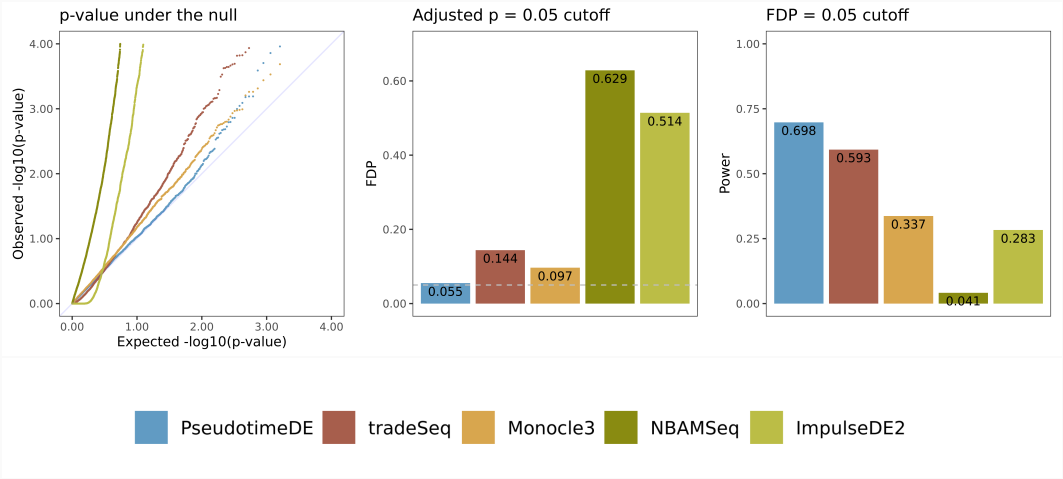
Discussion

- Clipper avoids the need for
  - valid (finite-sample or asymptotic) null distribution
  - high-resolution p-values
- Broad applications in high-throughput data analysis
  - key: **contrast scores**
- Importance of validating the FDR control
  - FDR control is bluntly assumed but rarely validated in most bioinformatics methods

Discussion

- Clipper avoids the need for
  - valid (finite-sample or asymptotic) null distribution
  - high-resolution p-values
- Broad applications in high-throughput data analysis
  - key: **contrast scores**
- Importance of validating the FDR control
  - FDR control is bluntly assumed but rarely validated in most bioinformatics methods
- Future work
  - incorporate feature covariates (e.g., gene variance)
  - choose contrast score & # of permutations: power
  - implement Clipper in bioinformatics tools

PseudotimeDE: Identification of DEGs along Pseudotime with Valid p-values



Dongyuan Song R package:

https://github.com/SONGDONGYUAN1994/PseudotimeDE

Acknowledgements

UCLA JSB:

- Dongyuan Song
- Elaine Huang
- Tianyi Sun
- Kexin Li
- Other lab members

City of Hope:

- Dr. Leo Wang
- MeiLu McDermott
- Kyla Woyshner
- Antigoni Manousopoulou

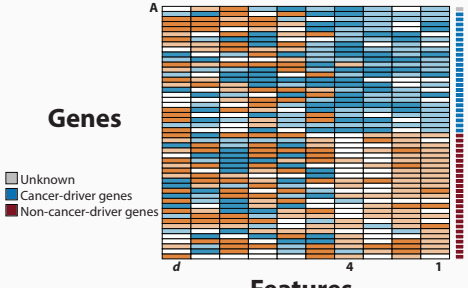
UCI

- Dr. Wei Li

Xinzhou Ge

Yiling Elaine Chen

Cancer-driver gene prediction: multiple testing?



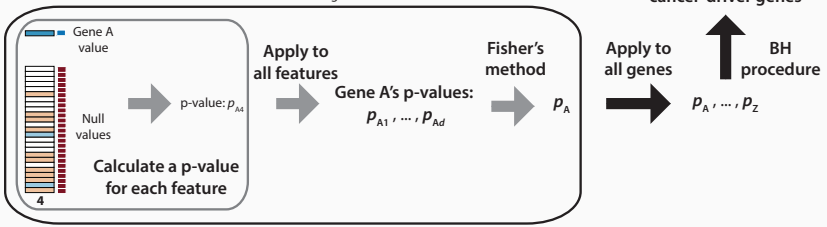
Cell

Cited for > 500 times

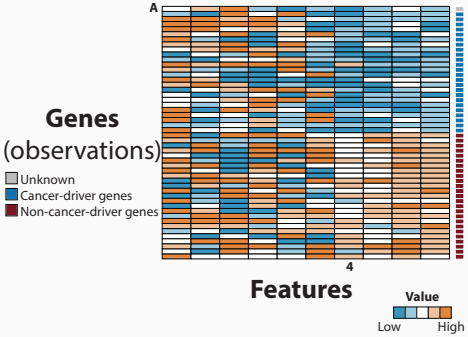
Resource

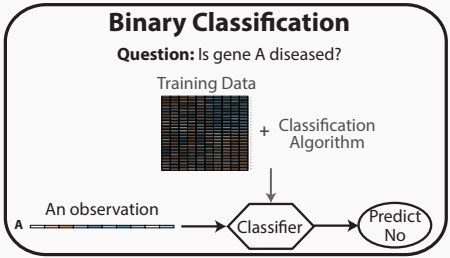
Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome

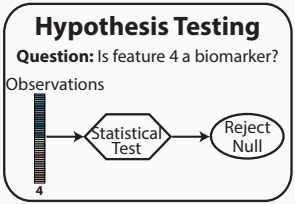
Teresa Davoli,[1,2,3] Andrew Wei Xu,[2,4,5] Kristen E. Mengwasser,[1,2] Laura M. Sack,[1,2] John C. Yoon,[2,3] Peter J. Park,[2,4] and Stephen J. Elledge[1,2,*]
[1]Howard Hughes Medical Institute, Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[2]Division of Genetics, Brigham and Women's Hospital, Boston, MA 02115, USA
[3]Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
[4]Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
[5]These authors contributed equally to this work
*Correspondence: selledge@genetics.med.harvard.edu
http://dx.doi.org/10.1016/j.cell.2013.10.011



Multiple testing vs. binary classification







SCIENCE ADVANCES | RESEARCH ARTICLE

CANCER

DORGE: Discovery of Oncogenes and tumoR suppressor genes using Genetic and Epigenetic features

Jie Lyu[1]*, Jingyi Jessica Li[2]*†, Jianzhong Su[3], Fanglue Peng[3], Yiling Elaine Chen[2], Xinzhou Ge[2], Wei Li[1]†

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