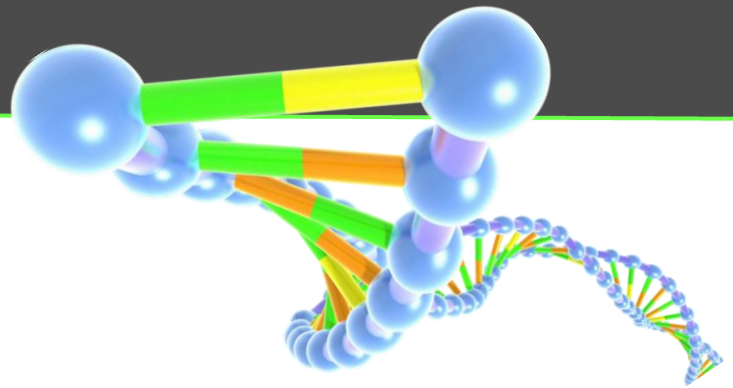


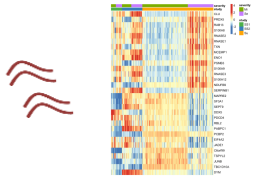
Identifying and Preventing Artifacts in High Dimensional Data: Computational Science in Immuno-Oncology

Hongkai Ji
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Email: hji@jhu.edu

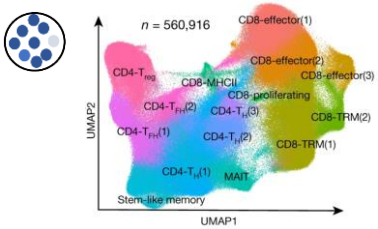
Background

Genome
(Whole-genome/exome sequencing)

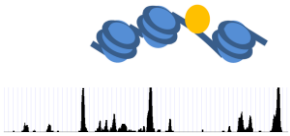
Transcriptome
(RNA-seq)

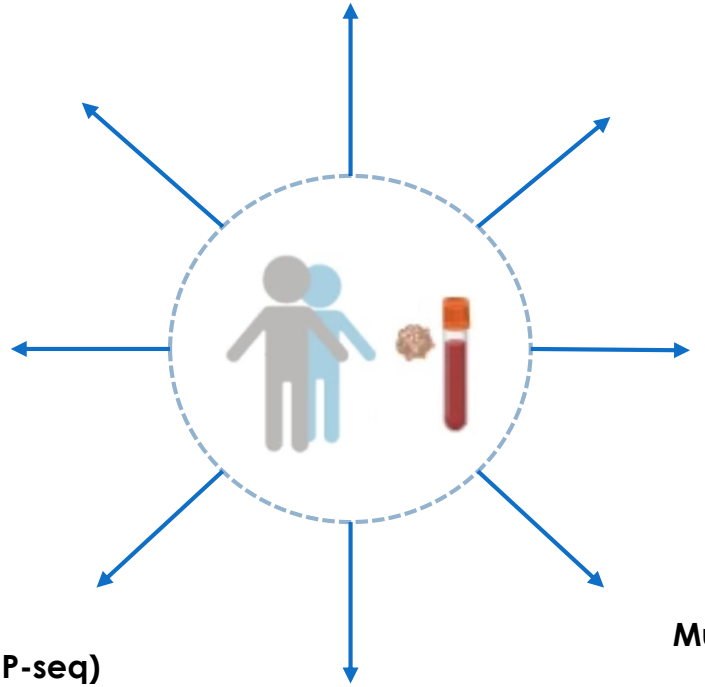


Single cell technologies

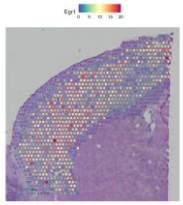


Regulome
(ATAC-seq/DNase-seq/ChIP-seq)

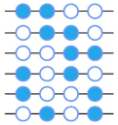




Spatial omics



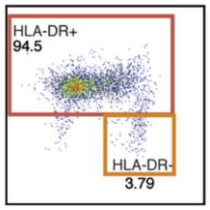
Epigenome
(Bisulfite Sequencing/ChIP-seq)



Immune repertoire
(TCR/BCR-seq)

CASSVEGGQGDTQYF
CASSQDWGDTEAFF
CASSVGDTEAFF
CASTMGDTYTDTQYF

Multicolor flow cytometry



Background

- Artifacts are common in data generated by high-throughput technologies

- Why do we care?
  - Identifying and preventing artifacts will help better discover true signals

Goal

- It is not our objective to provide an exhaustive list of all possible artifacts

Instead, we will

- Discuss common sources of artifacts
- Discuss some general principles and methods for identifying and preventing artifacts

Common sources of artifacts

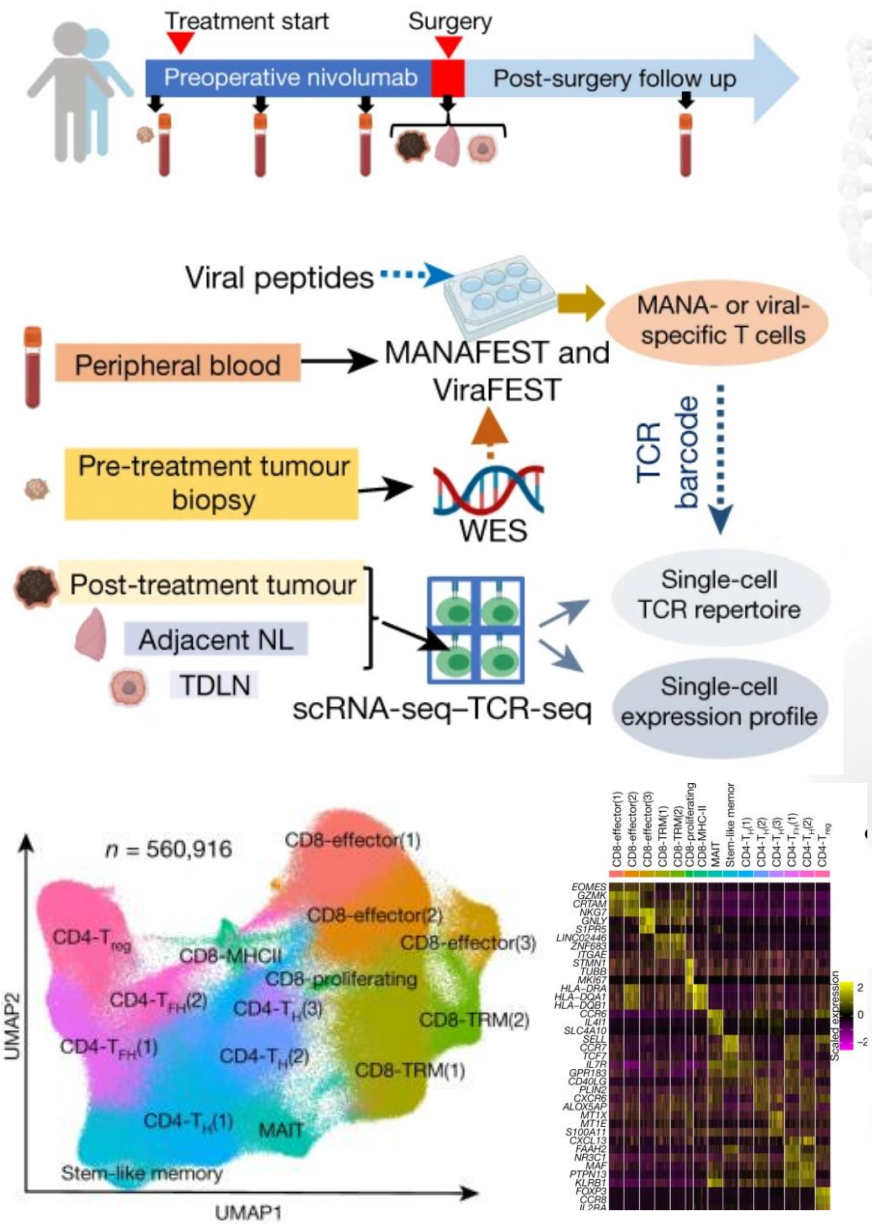


Image from Caushi et al. Nature. 596: 126–132, 2021

Study design

- Lack of proper control or randomization

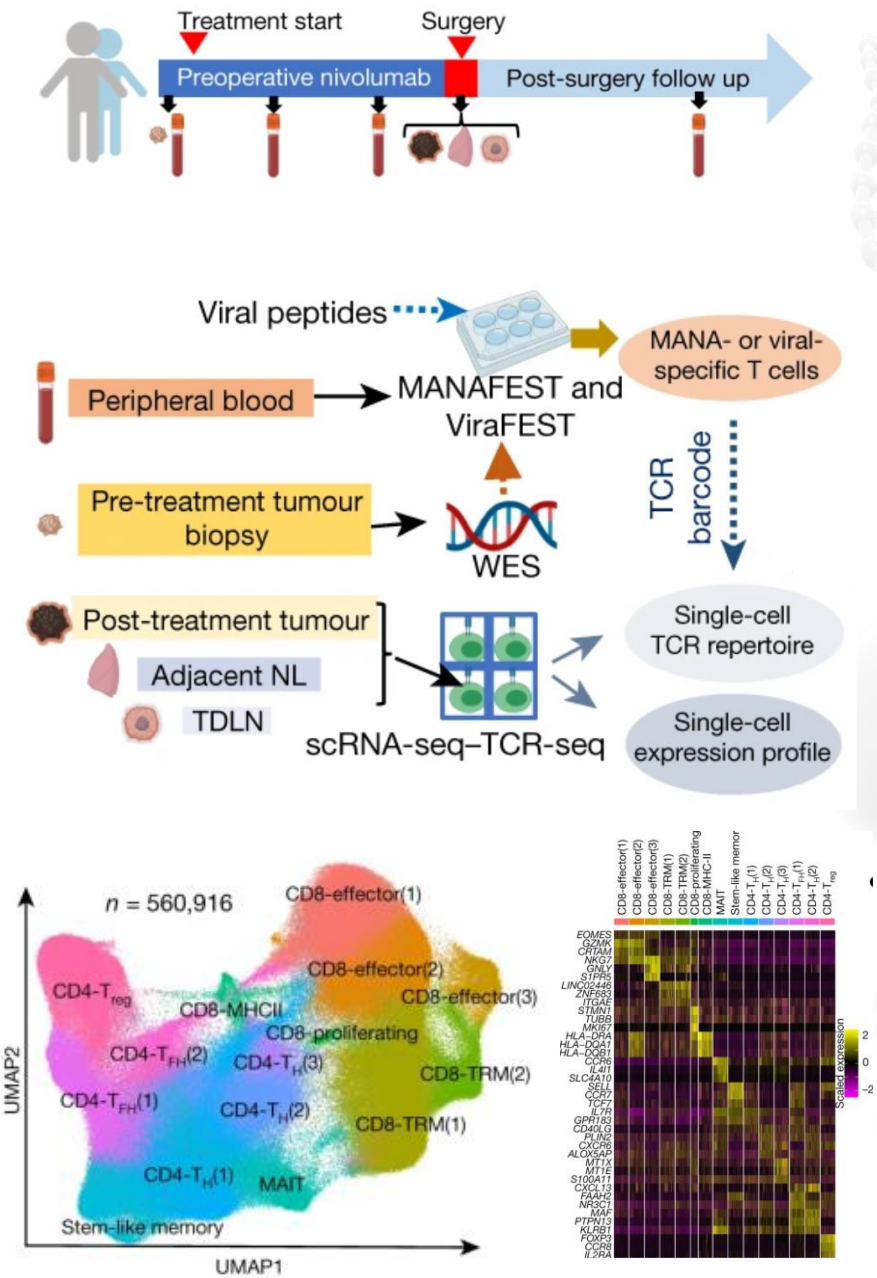
Data generation

- Bias and noise in technology
- Bias in experimental procedure

Data analysis

- Improper normalization
- Failure to control confounders
- Wrong models, assumptions or methods

Artifacts due to study design



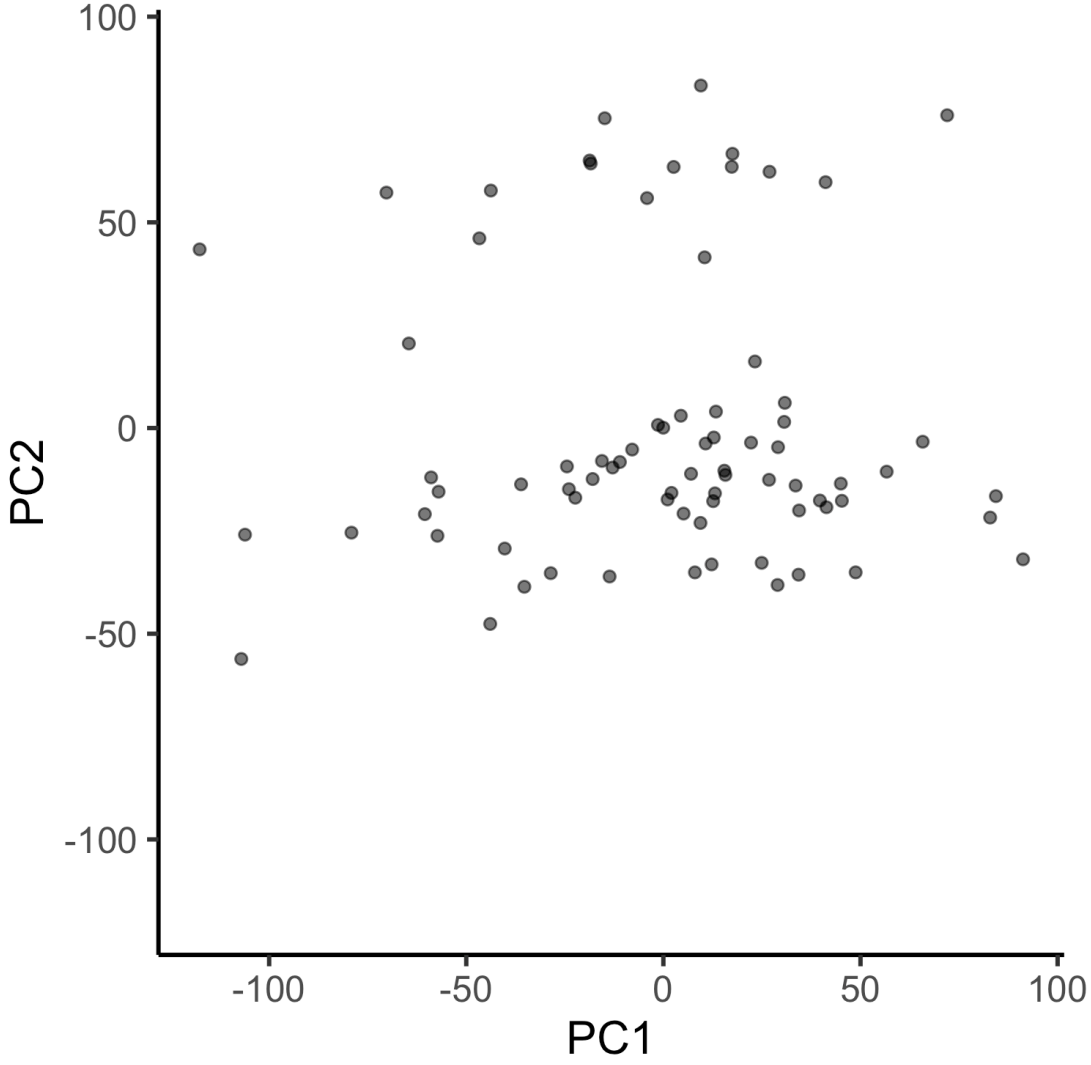
Study design

- Lack of proper control or randomization

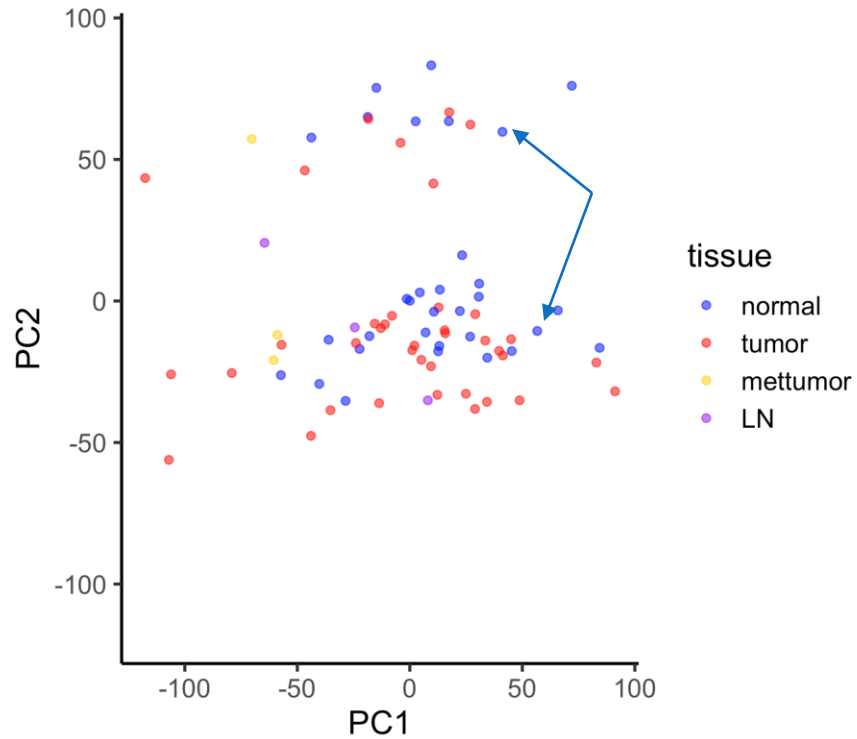
Data generation

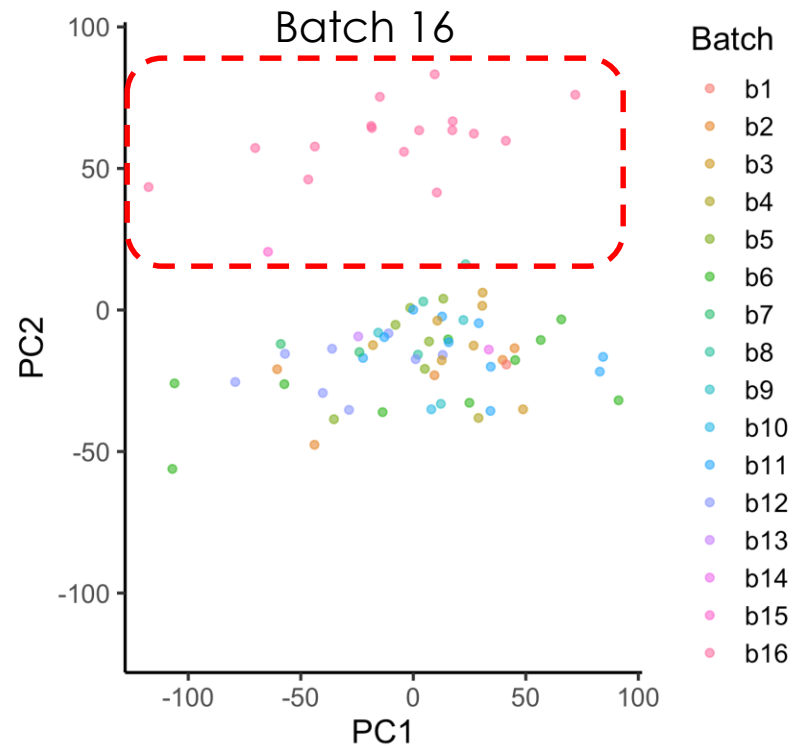
Data analysis

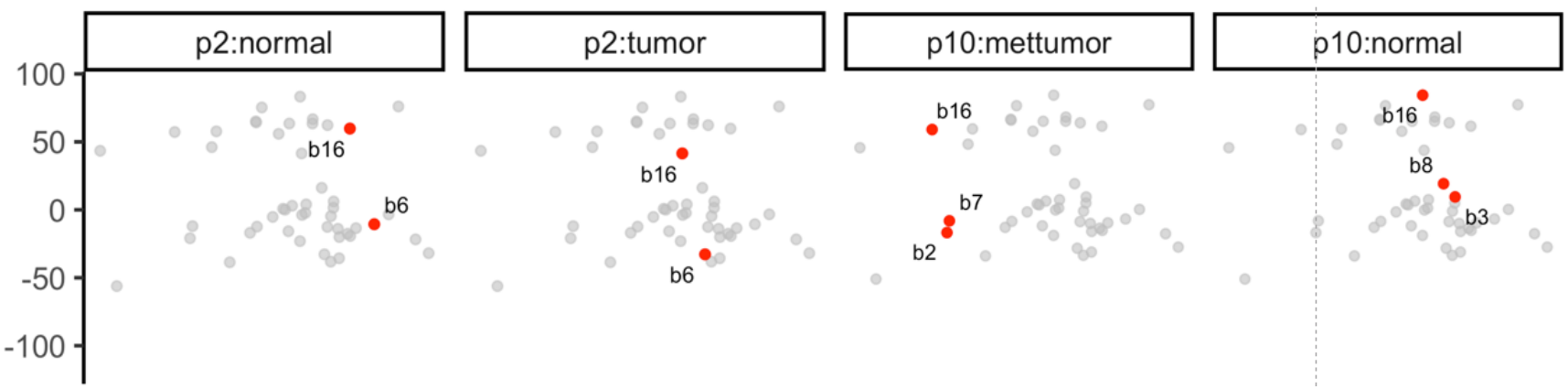
Example: Batch effects



Example: Batch effects



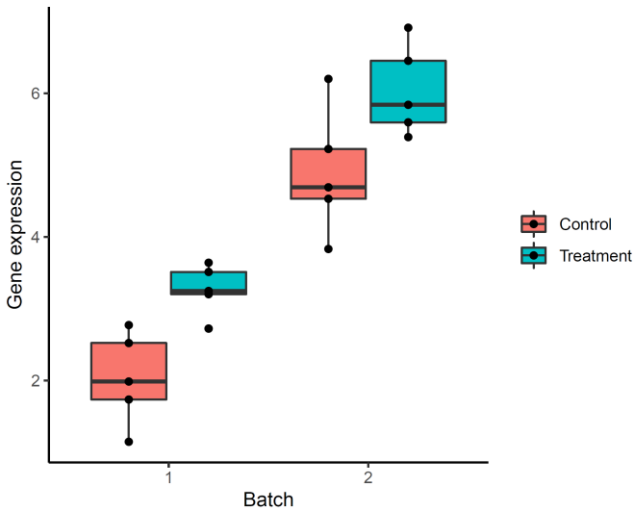




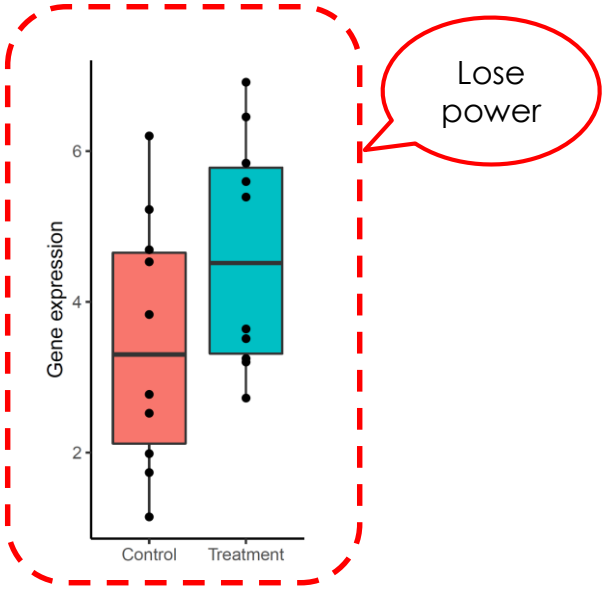
Batch effects: differential expression

A differential gene

Separate by batch



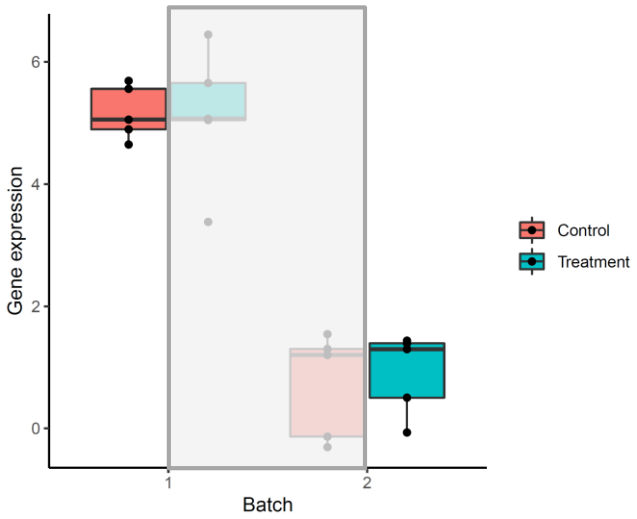
Ignore batch



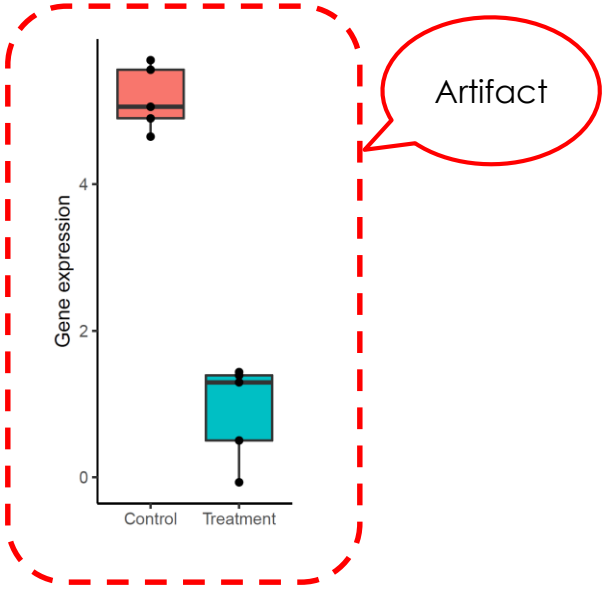
Batch effects: differential expression

A non-differential gene

Separate by batch



Treatment confounded with batch



Batch effects

PNAS

ARTICLES FRONT MATTER AUTHORS TOPICS +

LETTER | BIOLOGICAL SCIENCES | FREE ACCESS

Clarifying the effect of library batch on extracellular RNA sequencing

Christopher Hartl and Yuan Gao Authors Info & Affiliations

January 21, 2020 117(4)1849-1850 https://doi.org/10.1073/pnas.1916312117

Reanalysis of the raw data demonstrated a perfect confound between read length and cancer status (50 base pairs [bp] for both cancer cohorts, 75 bp for normal). Raw expression principal components PC1 and PC2, which separate cancer from normal samples, highly correlate to alignment metrics (Fig. 1 *A* and *B*). Following in silico read-length trimming, normal samples still exhibited perfect or near-perfect separation along a number of purely technical variables: mismatch rate, intronic rate, exonic rate, ribosomal RNA (rRNA) rate, and others (Fig. 1 *C* and *D*). Based on these observations, it seems that serum from individuals with cancer was processed separately from serum from individuals without cancer, creating a perfect confound between library batch, sequencing batch, and status. Since many standard RNA sequencing

How to prevent artifacts due to batches?

- Proper control and randomization

	T1	T2	C1	C2
Batch1	√	√		
Batch2			√	√

	T1	T2	C1	C2
Batch1	√		√	
Batch2		√		√

- For those who run experiments

  Team up with a statistician or experiment design expert before your study

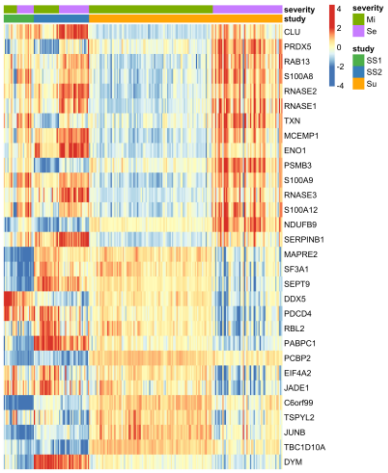
  Make sure everyone is on the same page

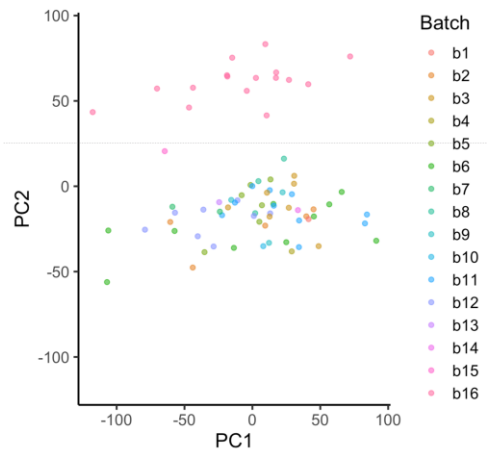
- For those who analyze data

  Talk to your wet lab collaborators before they generate data

How to identify batch artifacts?

- Review the study design
- Exploratory plots



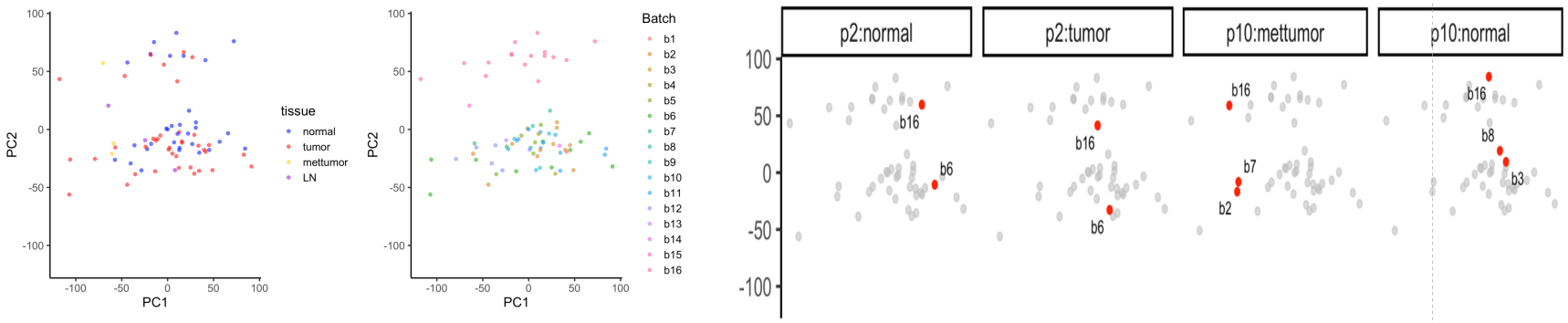


- Compare results from orthogonal datasets

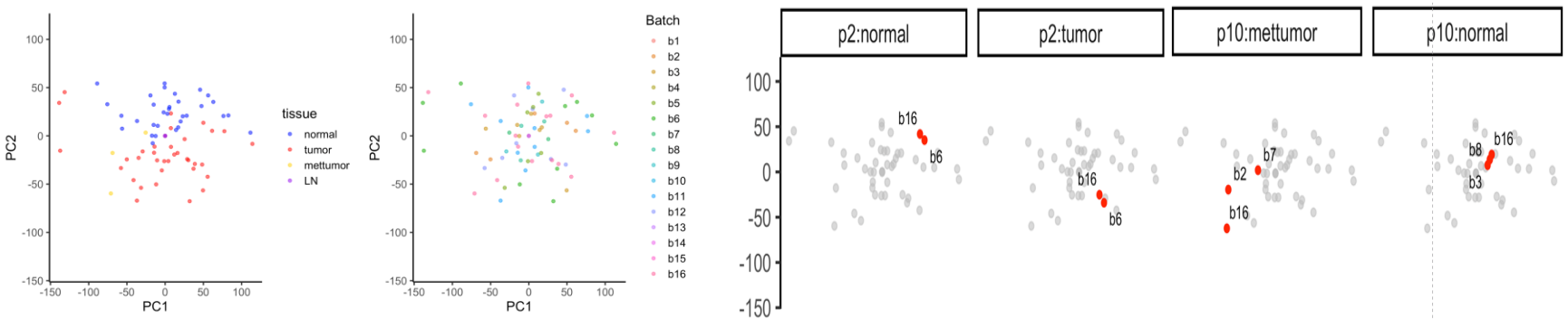
How to correct for batch effects?

- With proper design: regress out confounders

Before batch correction



After batch correction



How to correct for batch effects?

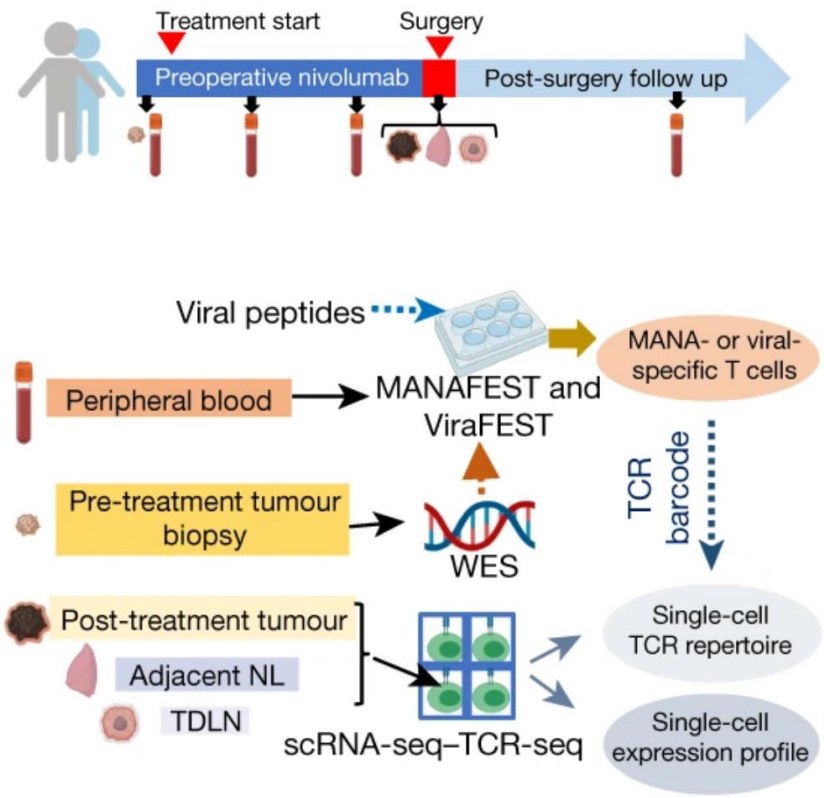
- With proper design: regress out confounders
  - ComBat (Johnson et al. Biostatistics, 8:118-127, 2007)
  - Surrogate variable analysis (Leek & Storey, PLoS Genet. 3:e161, 2007)
  - Remove unwanted variation (Gagnon-Bartsch & Speed, Biostatistics, 13:539-52, 2012)
  - A good review (Leek et al. Nat Rev Genet. 11: 733-739, 2010)

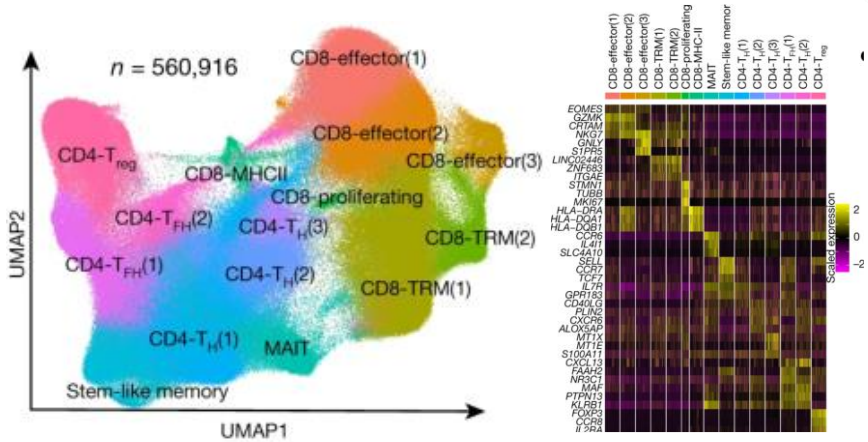
- With perfect confounding: profile new samples
  - Include samples to be compared in the same batch
  - Generate multiple batches to estimate batch variance

Remarks

- Batch effect is just one example of unwanted variation that may cause artifacts
- Other confounders may also create artifacts
- They often can be dealt with by following the same principles

Artifacts created during data generation





Study design

Data generation

- Bias and noise in technology
- Bias in experimental procedure

Data analysis

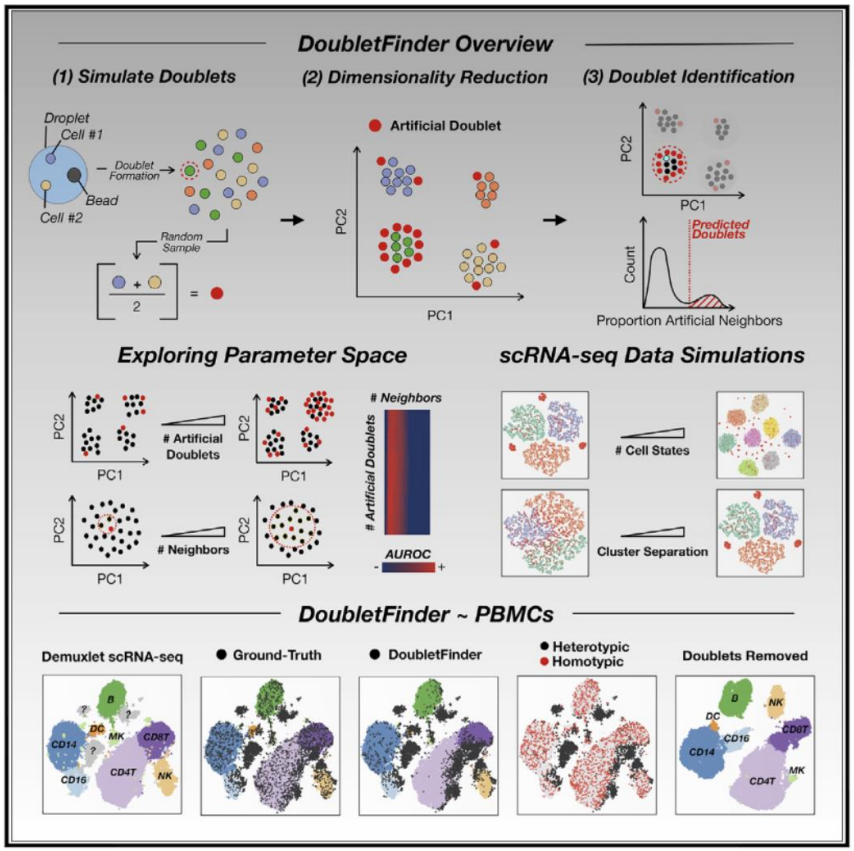
Example 1: Doublets in single-cell RNA-seq

Brief Report

Cell Systems

DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors

Graphical Abstract



Authors

Christopher S. McGinnis,
Lyndsay M. Murrow, Zev J. Gartner

Correspondence

zev.gartner@ucsf.edu

In Brief

scRNA-seq data interpretation is confounded by technical artifacts known as doublets—single-cell transcriptome data representing more than one cell. Moreover, scRNA-seq cellular throughput is purposefully limited to minimize doublet formation rates. By identifying cells sharing expression features with simulated doublets, DoubletFinder detects many real doublets and mitigates these two limitations.

McGinnis et al. Cell Systems. 8: 329–337, 2019

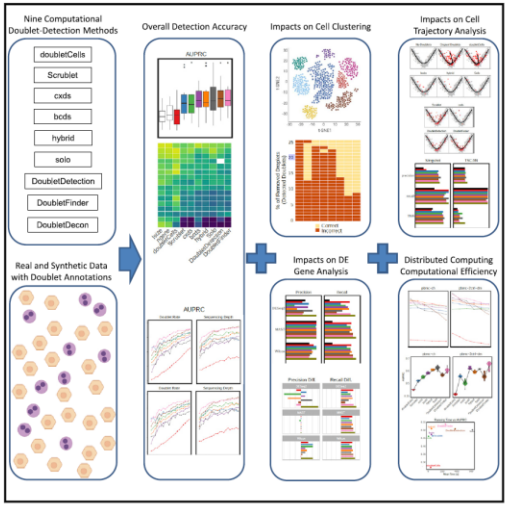
Example 1: Doublets in single-cell RNA-seq

Article

Cell Systems

Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data

Graphical Abstract



Authors

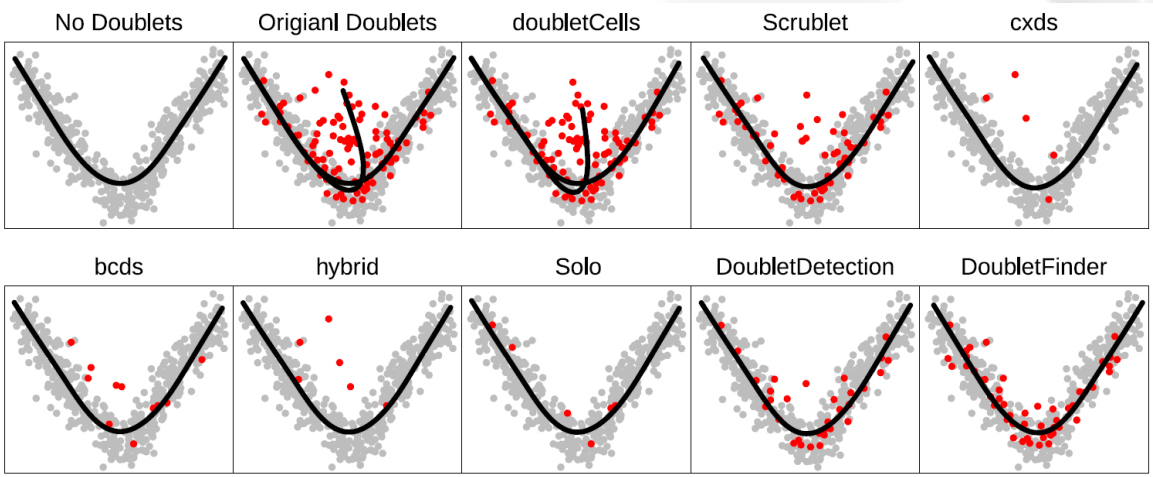
Nan Miles Xi, Jingyi Jessica Li

Correspondence

jli@stat.ucla.edu

In Brief

We conduct a systematic benchmark study of nine cutting-edge computational doublet-detection methods. We evaluate the methods' detection accuracy, impacts on downstream analyses, and computational efficiency, using a comprehensive set of real and synthetic data. Although no method dominates in all aspects, the DoubletFinder and cxds methods have the best detection accuracy and computational efficiency, respectively.



Xi & Li. Cell Systems. 12: 176–194, 2021

Example 2: Artifacts due to tissue dissociation

nature methods

Explore content ⌄ About the journal ⌄ Publish with us ⌄

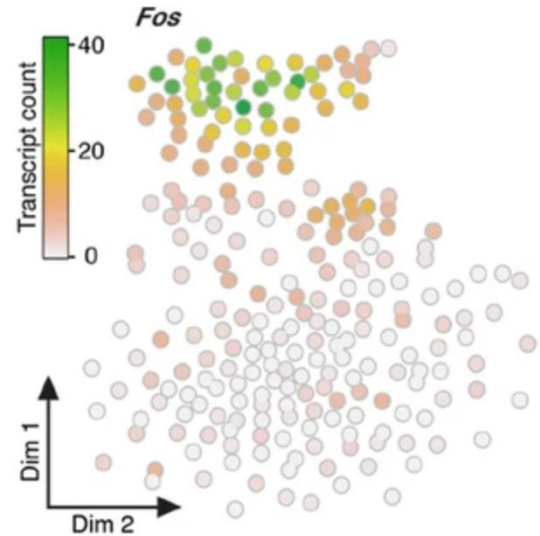
nature > nature methods > correspondence > article

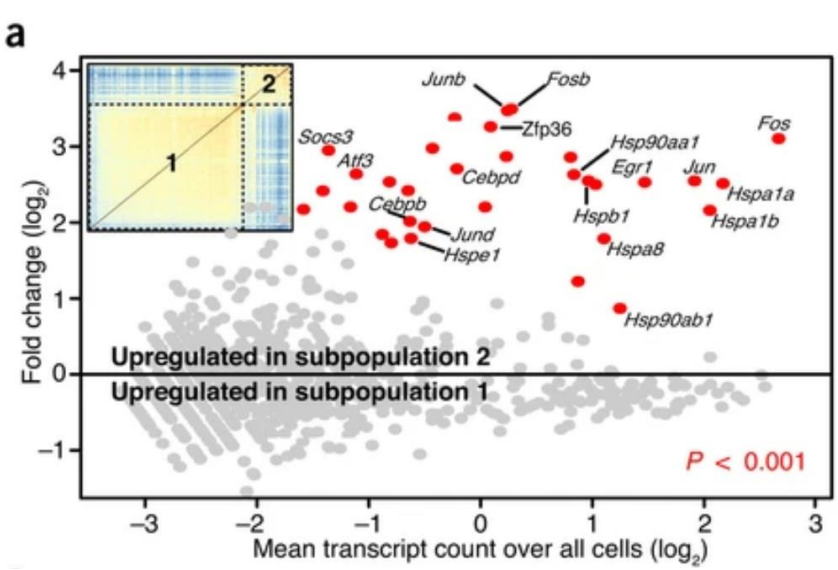
Published: 29 September 2017

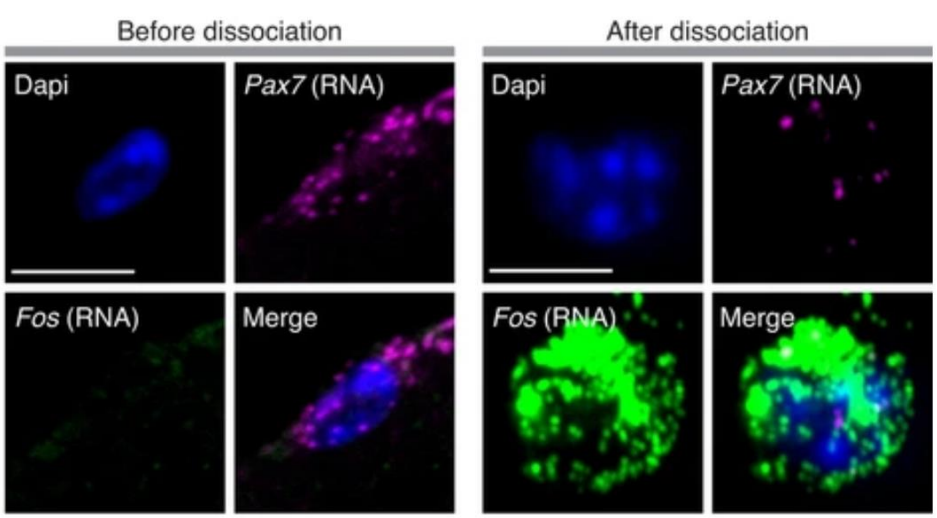
Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations

Susanne C van den Brink, Fanny Sage, Ábel Vértesy, Bastiaan Spanjaard, Josi Peterson-Maduro, Chloé S Baron, Catherine Robin & Alexander van Oudenaarden

Nature Methods **14**, 935–936 (2017) | Cite this article







Example 2: Artifacts due to tissue dissociation

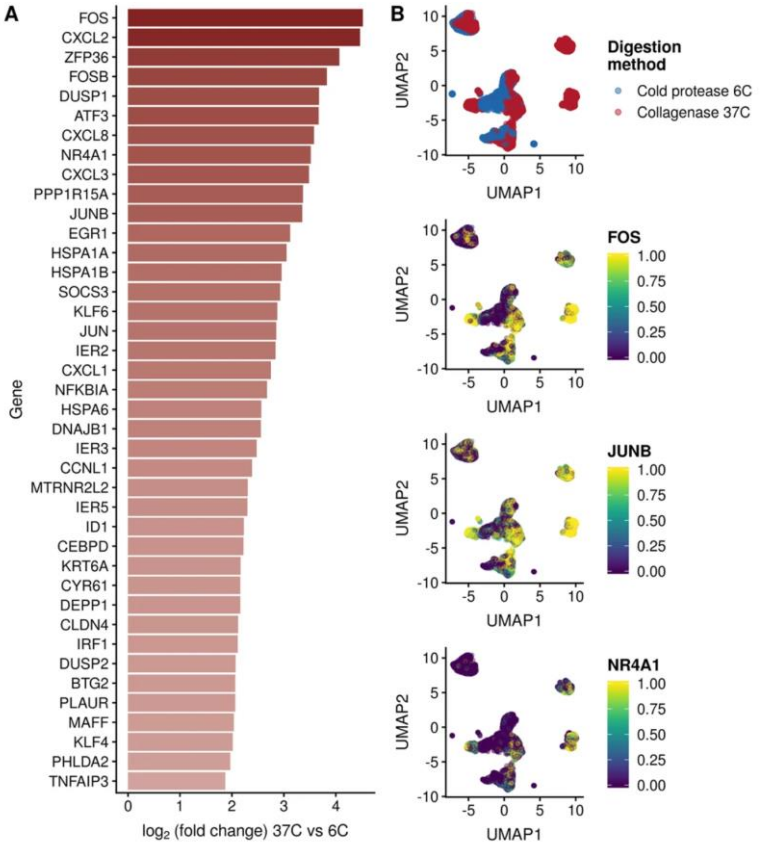
O'Flanagan *et al. Genome Biology* (2019) 20:210
https://doi.org/10.1186/s13059-019-1830-0

Genome Biology

RESEARCH Open Access

Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses

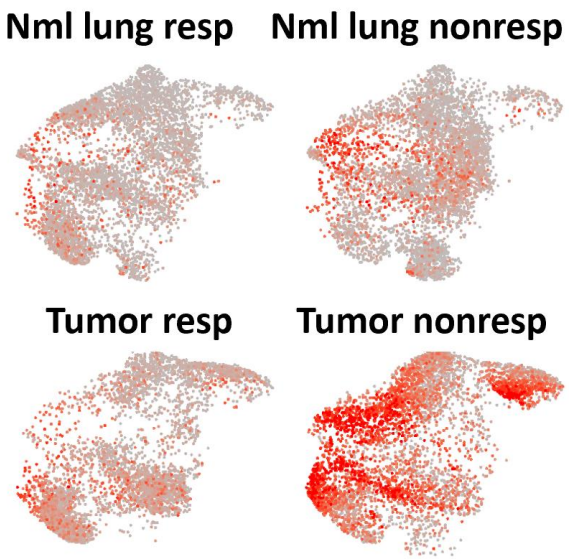
Ciara H. O'Flanagan[1†], Kieran R. Campbell[1,2,3†], Allen W. Zhang[1,4,5†], Farhia Kabeer[1,6†], Jamie L. P. Lim[7], Justina Biele[1], Peter Eirew[1], Daniel Lai[1], Andrew McPherson[1,7], Esther Kong[1], Cherie Bates[1], Kelly Borkowski[1], Matt Wiens[1], Brittany Hewitson[1], James Hopkins[1], Jenifer Pham[1], Nicholas Ceglia[4], Richard Moore[8], Andrew J. Mungall[8], Jessica N. McAlpine[9], The CRUK IMAXT Grand Challenge Team[1], Sohrab P. Shah[1,6,7*] and Samuel Aparicio[1,3*]

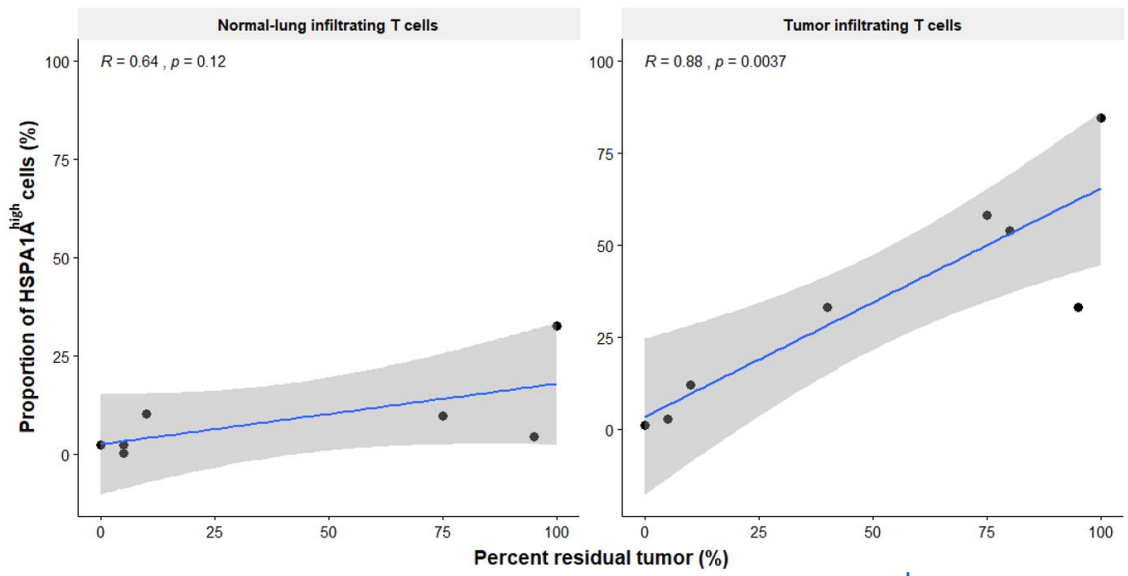


Dissociation using collagenase at 37°C results in a stress response as compared to dissociation using a cold active protease at 6°C.

Example 2: Artifacts due to tissue dissociation

HSPA1A expression



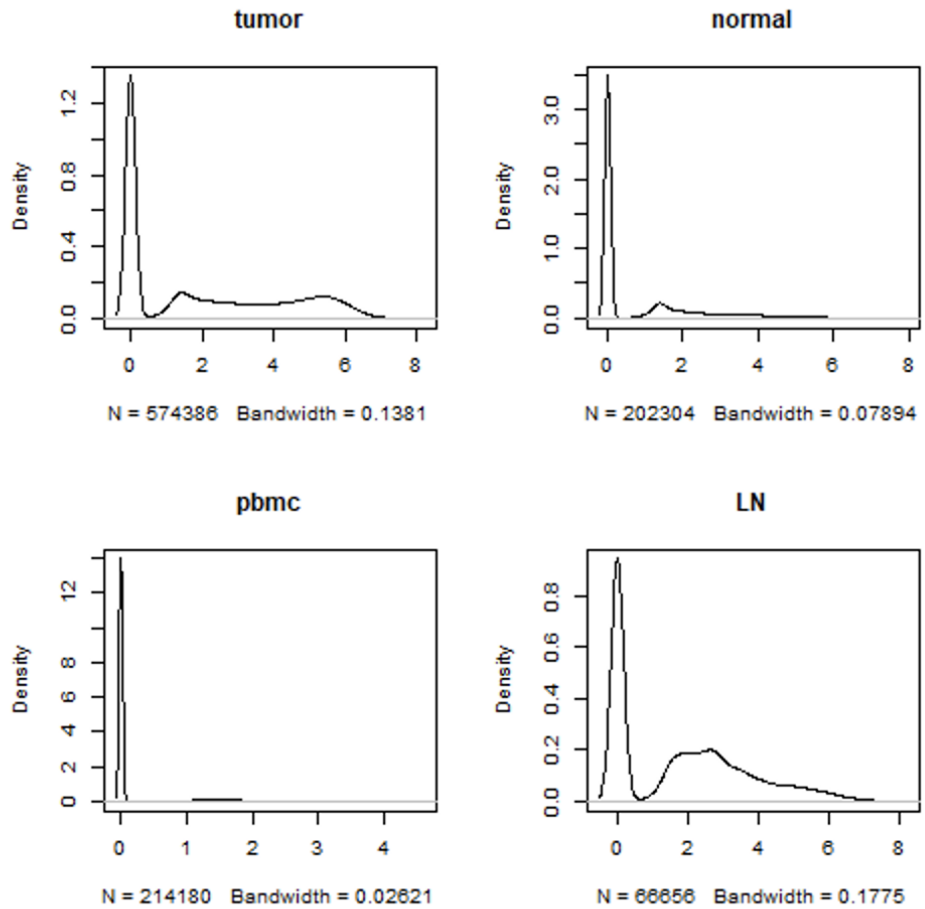


The correlation is absent in immunohistochemistry staining

Zhang, Caushi, Pardoll, Smith, et al.

Example 2: Artifacts due to tissue dissociation

HSPA1A only expresses in solid tissue samples but not in PBMC
(PBMC is handled without using dissociation enzyme)



Zhang, Caushi, Pardoll, Smith, et al.

How to deal with artifacts due to data generation?

Build knowledge

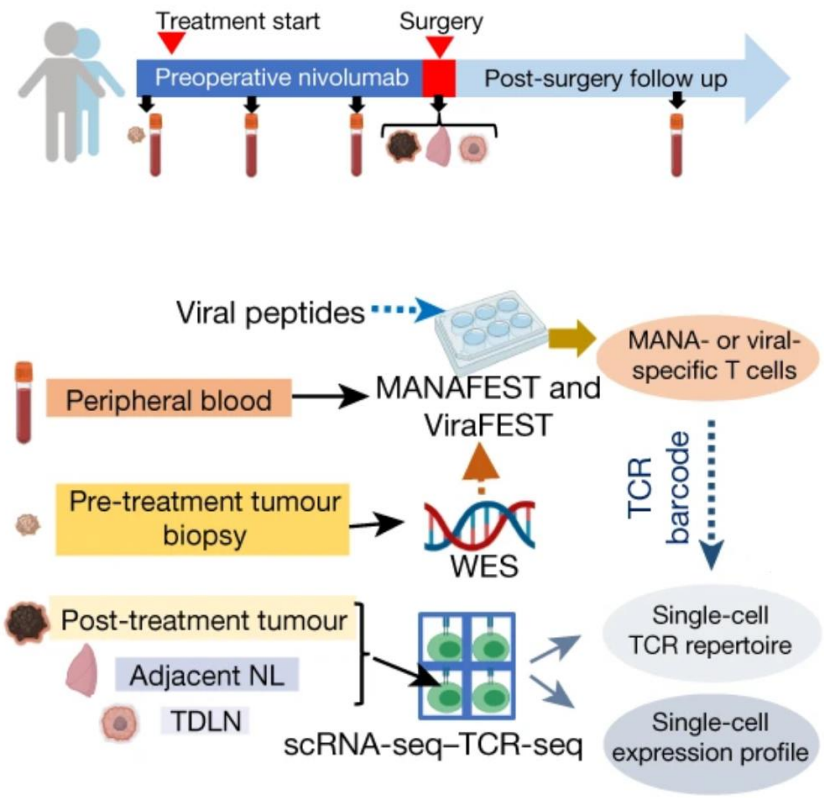
- Understand the data generation process
- Compare results from orthogonal data
- Analyze many datasets and find recurring patterns

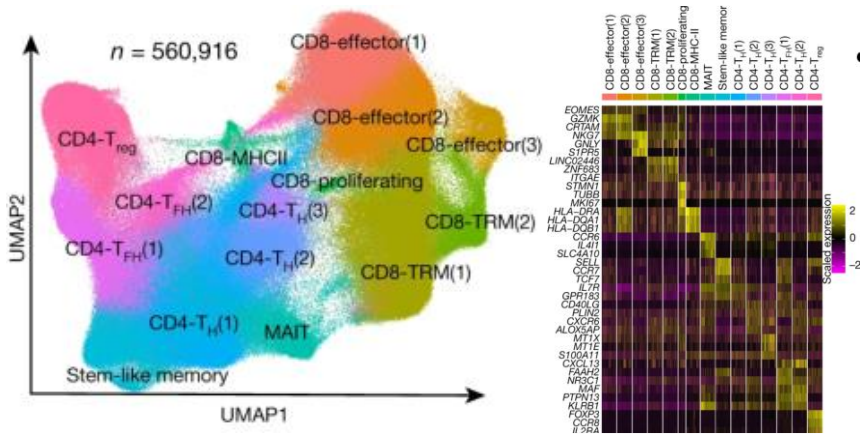
Develop solution

- Improve experimental technologies
- Develop computational algorithms for artifact detection and removal
- Spike-in experiments for benchmark

Importantly, wet lab and dry lab investigators need to closely work together!

Artifacts due to data analysis





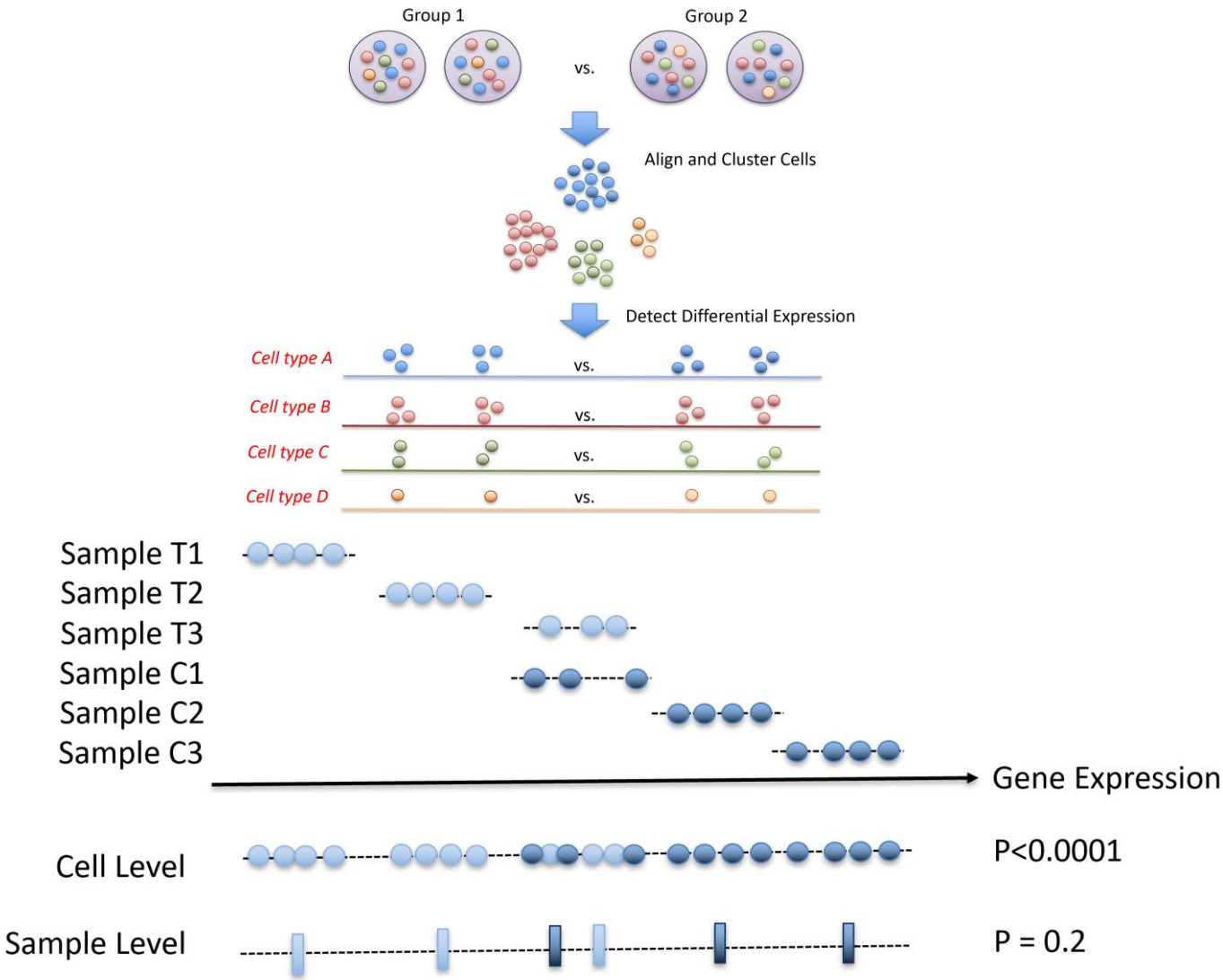
Study design

Data generation

Data analysis

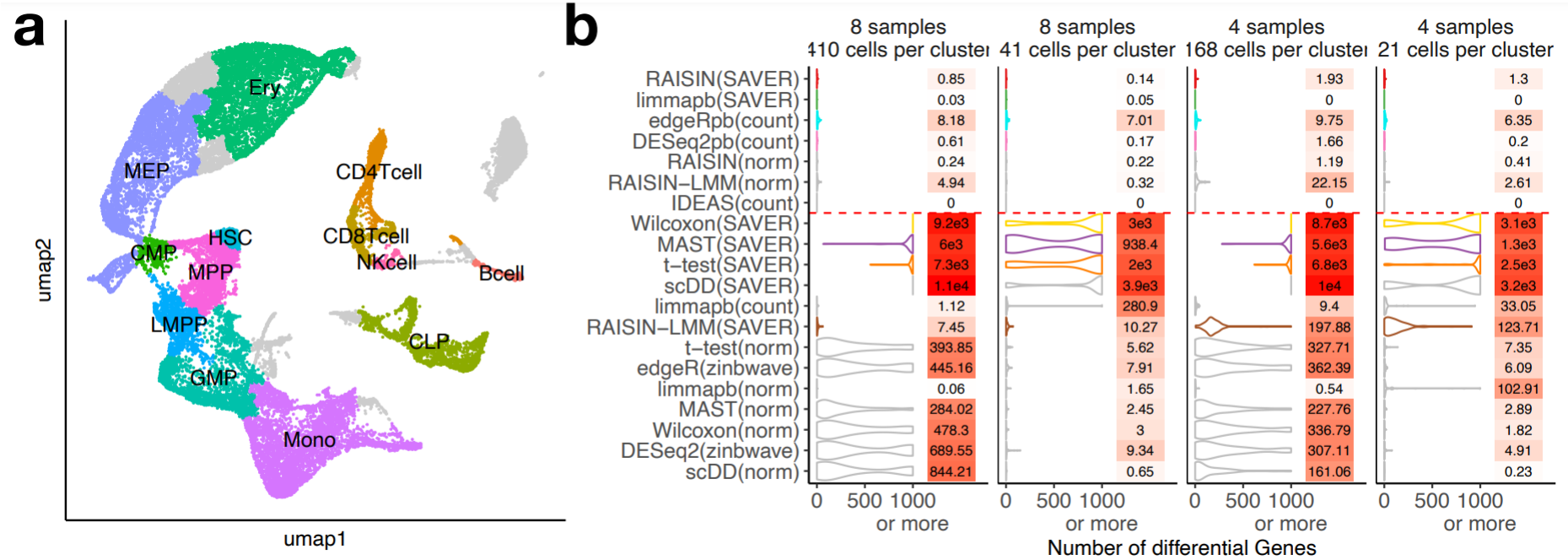
- Improper normalization
- Failure to control confounders
- Wrong models, assumptions or methods

Example 1: scRNA-seq differential expression

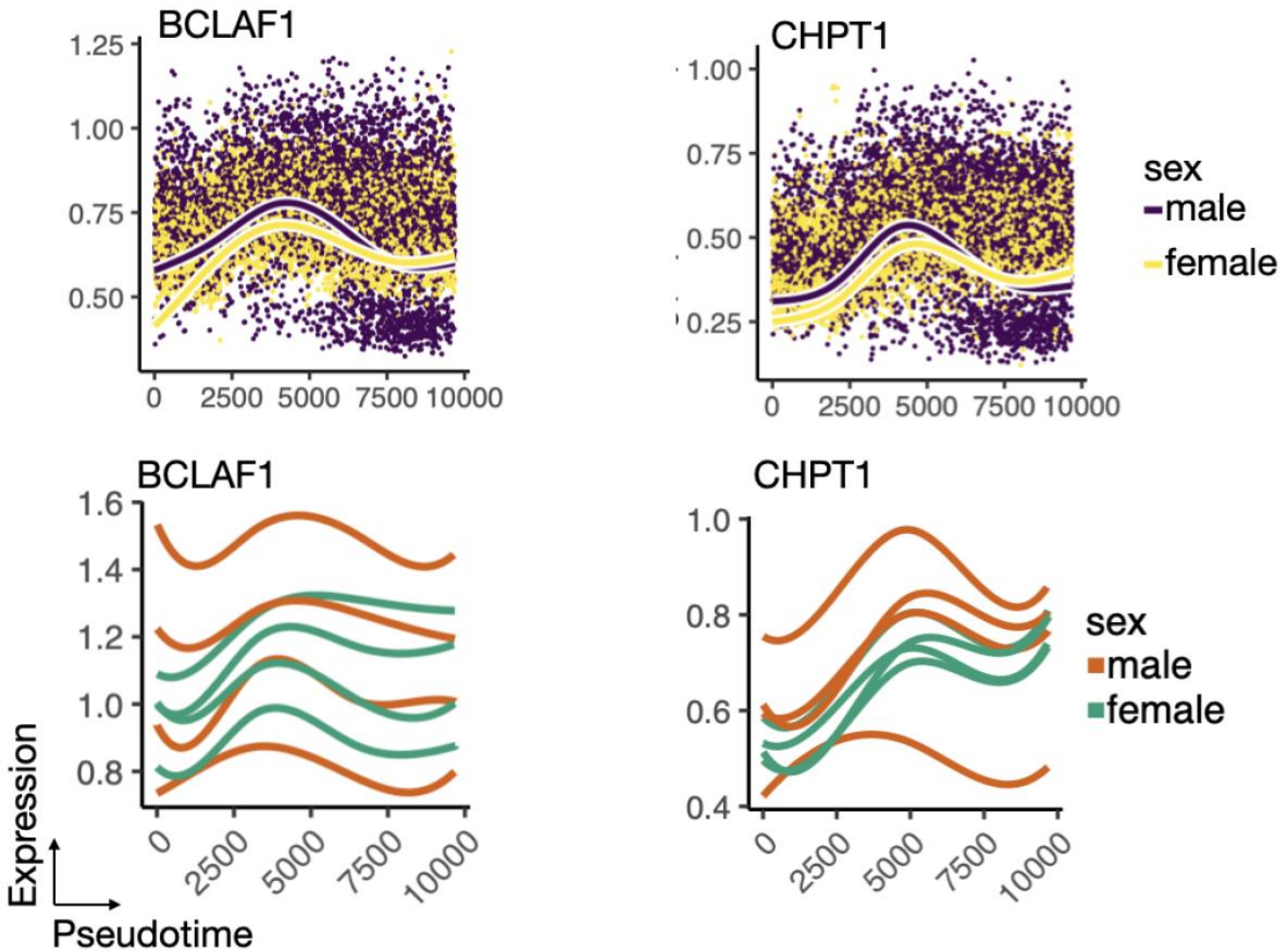


Wilcoxon test ignores sample variability.
When there are multiple samples, it will create false discoveries.

Example 1: scRNA-seq differential expression



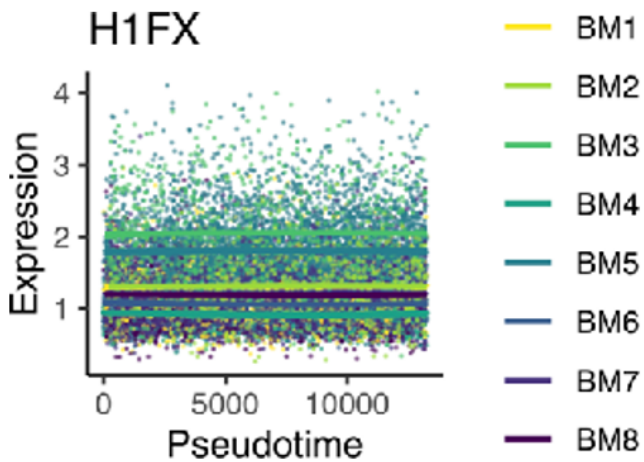
Example 1: scRNA-seq differential pseudotemporal expression



Hou et al. bioRxiv preprint, https://doi.org/10.1101/2021.07.10.451910

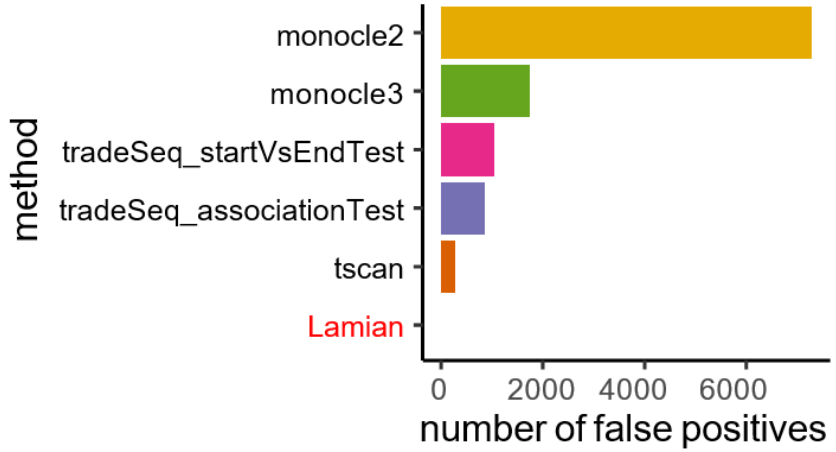
Example 1: scRNA-seq differential pseudotemporal expression

a



Method	FDR
tradeSeq_startVsEndTest	1.88×10^{-4}
tradeSeq_associationTest	6.09×10^{-6}
Monocle2	4.13×10^{-10}
Monocle3	1.70×10^{-3}
TSCAN	1.68×10^{-2}
Lamian	0.999

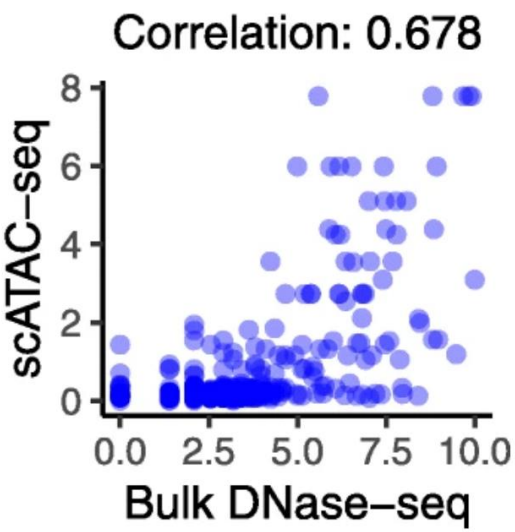
b



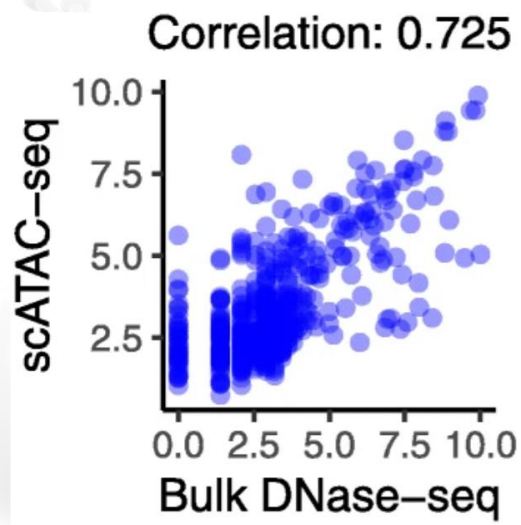
Hou et al. bioRxiv preprint, https://doi.org/10.1101/2021.07.10.451910

Example 2: Chromatin accessibility locus effects

GM12878 scATAC-seq vs. GM12878 bulk DNase-seq

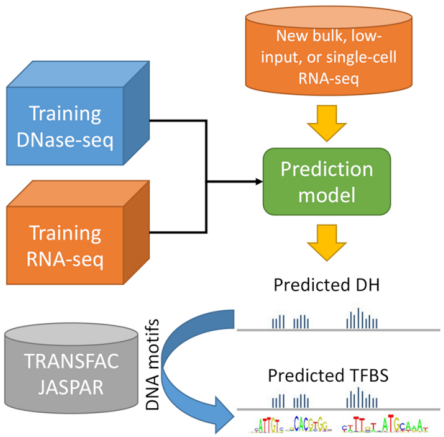


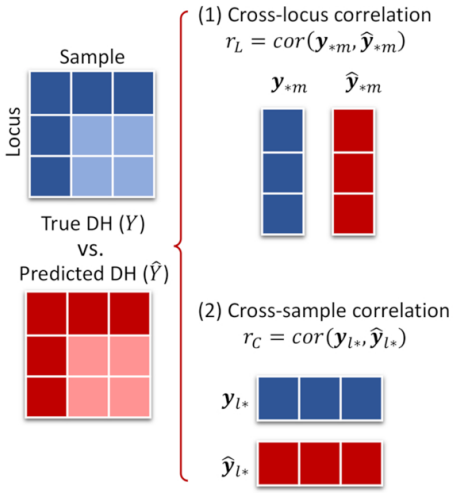
GM12878 scATAC-seq vs. ENCODE average bulk DNase-seq

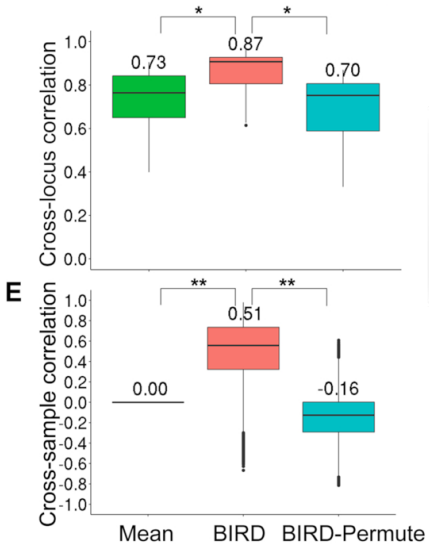


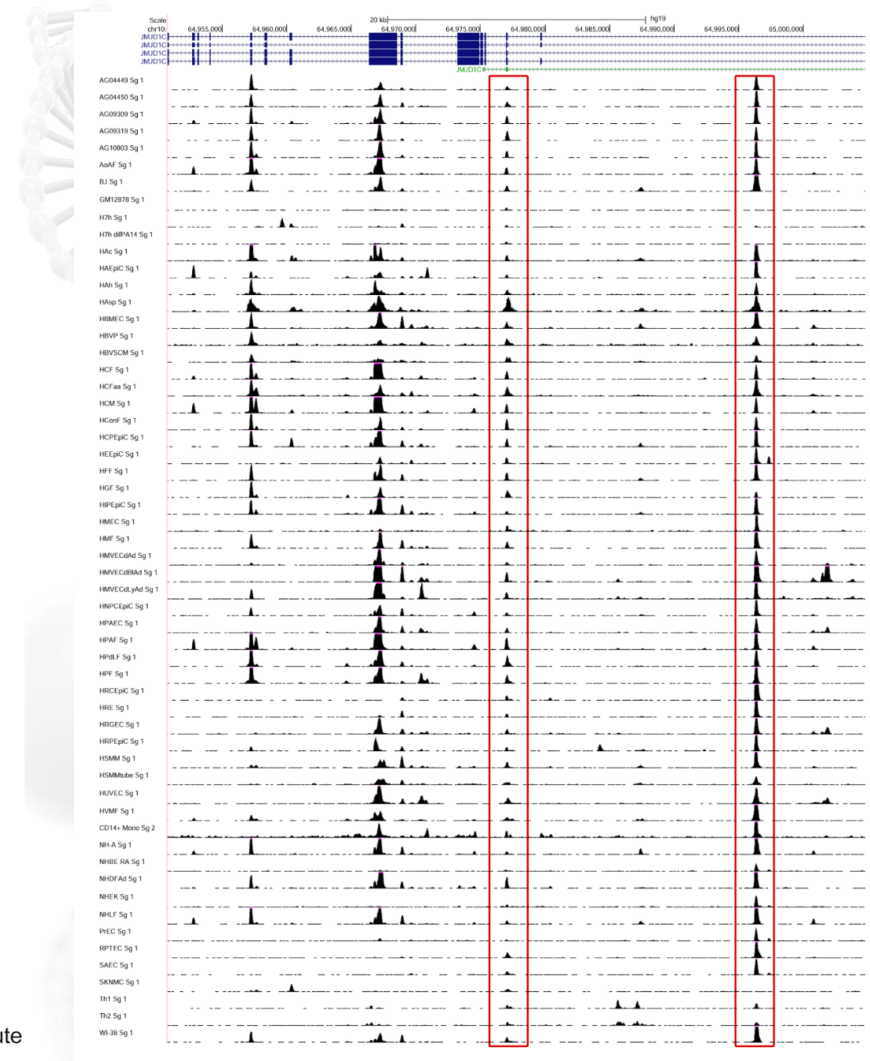
Ji et al. Genome Biology. 21: 161, 2020

Example 2: Chromatin accessibility locus effects







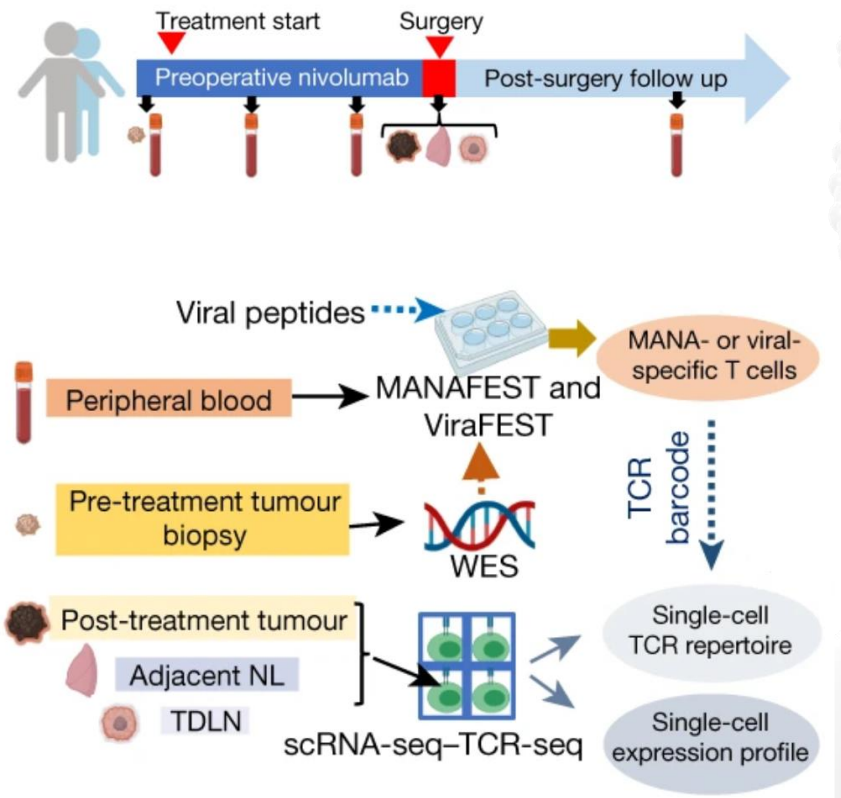


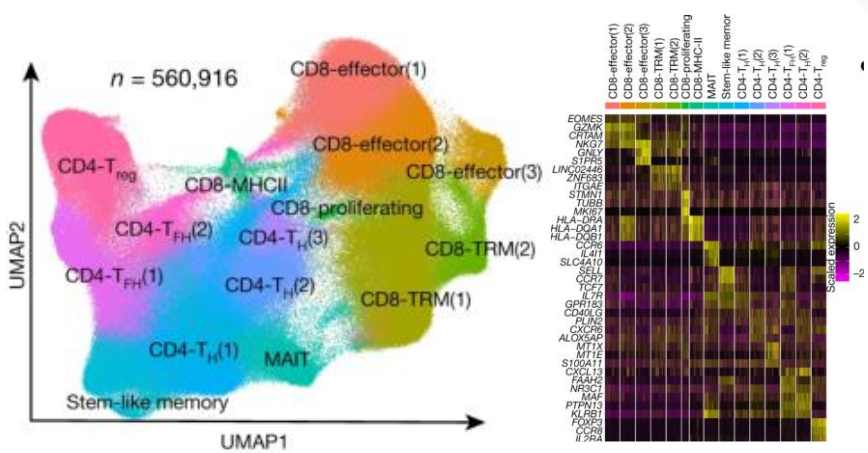
Zhou et al. Nature Communications. 8: 1038, 2017
Zhou et al. Nucleic Acids Res. 47: e121, 2019

How to prevent and identify artifacts due to data analysis?

- Build good understanding of your data
- Choose appropriate models and assumptions
- Use proper controls
- Learn from analyzing many datasets
- Compare results from orthogonal data
- Benchmark using spike-in experiments
- Again, wet lab and dry lab investigators should work closely together

Summary: common sources of artifacts





Study design

- Lack of proper control or randomization

Data generation

- Bias and noise in technology
- Bias in experimental procedure

Data analysis

- Improper normalization
- Failure to control confounders
- Wrong models, assumptions or methods

General principles and common methods for preventing, identifying and removing artifacts

- Wet lab and dry lab investigators work closely together from day one
- Use proper control and randomization
- Validate findings using orthogonal data
- Learn from analyzing many datasets
- Use appropriate models, assumptions, and analysis methods
- Benchmark using spike-ins

Acknowledgment

The Johns Hopkins Bloomberg School of Public Health

Boyang Zhang
Weiqiang Zhou
Ruzhang Zhao
Wenpin Hou
Stephanie Hicks

The Johns Hopkins School of Medicine

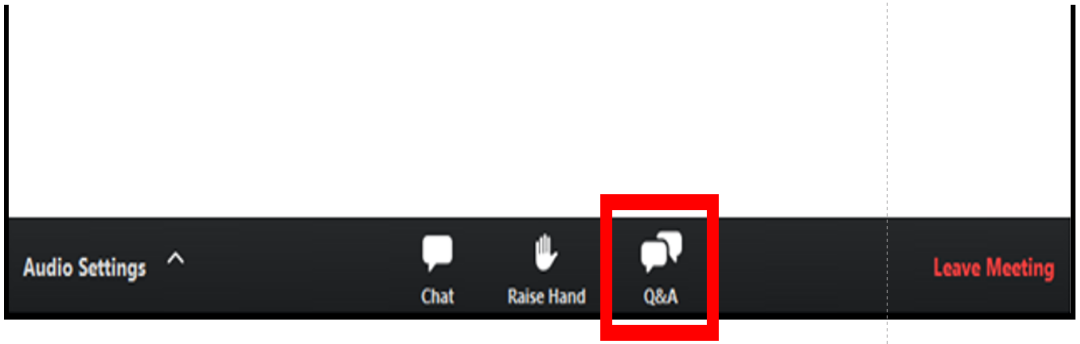
Jiajia Zhang
Justina X. Caushi
Arbor G. Dykema
Srinivasan Yegnasubramanian
Drew M. Pardoll
Kellie N. Smith

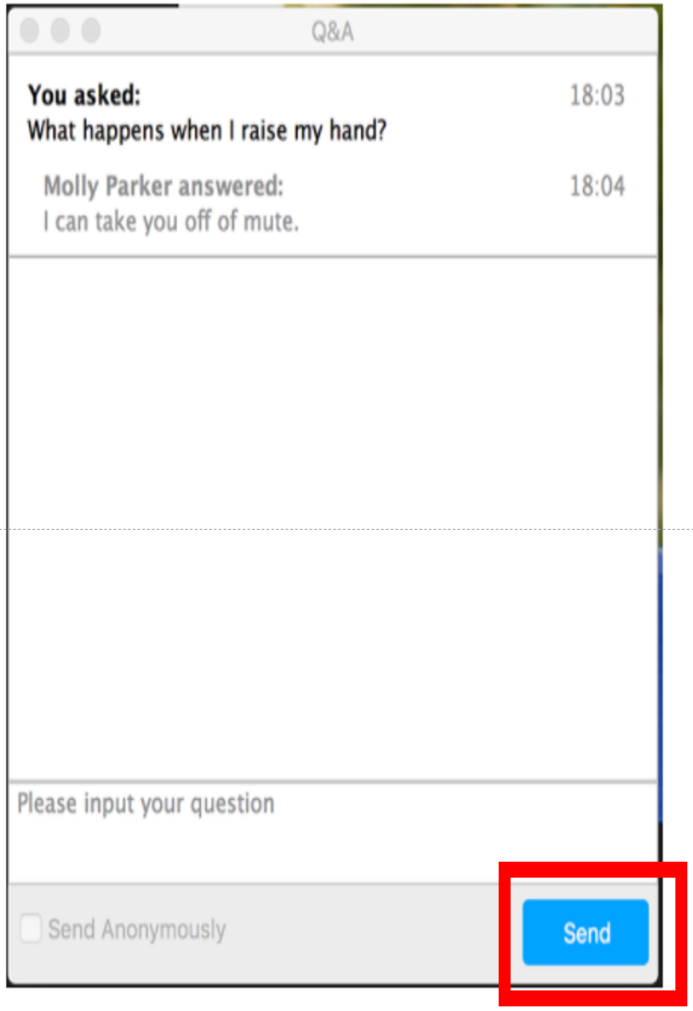
Duke University School of Medicine
Zhicheng Ji

Funding
NIH R01HG009518, R01HG010889
Johns Hopkins IDIES Seed Fund

How to Submit Questions

- Click the "Q&A" icon located on at the bottom of your Zoom control panel
- Type your question in the Q&A box, then click "Send"
- Questions will be answered in the Question & Answer session at the end of the webinar (as time permits)





SITC-NCI Computational Immuno-oncology
WEBINAR SERIES