Promoting Statistical Rigor and Reproducibility: Awareness, Action, and Advocacy

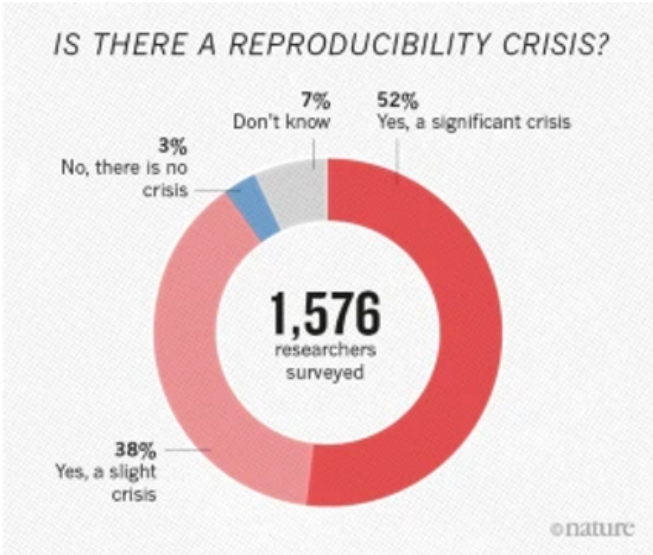
Zhaoxia Yu

Professor, Department of Statistics, University of California, Irvine

UCI, Feburary 2025

Reproducibility/Replication

- A cornerstone of scientific research
- An existing crisis is reproducibility/replication crisis
- Consequences of reproducibility crisis: wasting public resources, loosing public confidence, threatening people's health



Nature **volume 533**, pages452–454 (2016)

Examples

Science

Current Issue | First release papers | Archive | About | Sub

HOME > SCIENCE > VOL. 346, NO. 6216 > FIXING PROBLEMS WITH CELL LINES

POLICY FORUM | CELL BIOLOGY

Fixing problems with cell lines

Technologies and policies can improve authentication

JON R. LORSCH , FRANCIS S. COLLINS, AND JENNIFER LIPPINCOTT-SCHWARTZ Authors Info & Affiliations

SCIENCE · 19 Dec 2014 · Vol 346, Issue 6216 · pp. 1452-1453 · DOI: 10.1126/science.1259110

736 135

ILLUSTRATION: PETER AND MARIA HOEY/WWW.PETERHOEY.COM

Since the 1960's, more than 400 widely used cell lines worldwide have been shown to have been misidentified.

Studies using just two misidentified cell lines were included in three grants funded by NIH, two critical trials, 11 patents, and >100 papers.

nature Vol 440|16 March 2006|doi:10.1038/nature04533

LETTERS

A specific amyloid-β protein assembly in the brain impairs memory

Sylvain Lesné[1], Ming Teng Koh[4], Linda Kotilinek[1], Rakez Kayed[6], Charles G. Glabe[6], Austin Yang[7], Michela Gallagher[4] & Karen H. Ashe[1,2,3,5]

Memory function often declines with age[1], and is believed to deteriorate initially because of changes in synaptic function rather than loss of neurons[2]. Some individuals then go on to develop Alzheimer's disease with neurodegeneration. Here we use *Tg2576* mice, which express a human amyloid-β precursor protein (APP) variant linked to Alzheimer's disease, to investigate the cause of memory decline in the absence of neurodegeneration or amyloid-β protein amyloidosis. Young *Tg2576* mice (<6 months old) have normal memory and lack neuropathology, middle-aged mice (6–14 months old) develop memory deficits without neuronal loss, and old mice (>14 months old) form abundant neuritic plaques containing amyloid-β (refs 3–6). We found that memory deficits in middle-aged *Tg2576* mice are caused by the extracellular accumulation of a 56-kDa soluble amyloid-β assembly, which we term Aβ*56 (Aβ star 56). Aβ*56 purified from the brains of impaired *Tg2576* mice disrupts memory when administered to young rats. We propose that Aβ*56 impairs memory independently of plaques or neuronal loss, and may contribute to cognitive deficits associated with Alzheimer's disease.

Poor memory function can predict Alzheimer's disease [illegible] to 15 years before diagnosis[7], and non-demented individuals at risk genetically for Alzheimer's disease show abnormalities in functional brain imaging tests[8,9]. These and other studies imply that Alzheimer's disease has an insidious onset, which blurs the boundary between age-associated memory impairment and Alzheimer's disease[10]. *Tg(APPSWE)2576Kahs* mice (hereafter *Tg2576* mice) express a human amyloid-β precursor protein (APP) variant linked to Alzheimer's disease, and develop [illegible] neuropathological features of Alzheimer's, including amyloid plaques, dystrophic neurites and inflammatory changes[11]. However, *Tg2576* mice lack neurofibrillary tangles, significant neuronal loss and gross atrophy[4]. They may therefore be a good model to study pre-clinical stages of Alzheimer's disease, possible diagnosis of dementia or the onset of neuronal loss.

In *Tg2576* mice, as in other APP transgenic mice, there is strong evidence that [illegible] (Aβ) is responsible for age-related memory [illegible]

for memory loss, is associated with no change in memory function. One solution to this conundrum is to posit the existence of soluble Aβ assemblies that disrupt memory[4,5] which we designated Aβ* (Aβ star) and sought to identify in *Tg2576* mice.

A challenge in analysing Aβ in the brain lies in reliably separating the specific cellular pool of Aβ (for example, extracellular, intracellular, membrane-bound and insoluble). We overcame this obstacle by developing a high-fidelity extraction procedure that separates proteins in known cellular compartments (Supplementary Fig. 1). Our new extraction method allowed us to quantify and compare four independent pools of transgene-derived Aβ species.

To resolve the problem of a mismatch between Aβ levels and memory deficits, we used our extraction procedure to search for Aβ* in *Tg2576* mice between 4 and 25 months of age. We required candidate Aβ* molecules to satisfy two criteria. First, their appearance should coincide with memory loss at 6 months. Second, their levels should remain stable in middle-aged mice (6–14 months old). By immunoblotting immunoglobulin-depleted forebrain extracts, we found a set of apparent assemblies of Aβ in the soluble, extracellular-enriched fraction from 6-month-old mice (Fig. 1c). In addition to a faint 4-kDa band corresponding to Aβ monomers, 6E10- and 4G8-immunoreactive proteins (see Methods) were detected at molecular masses theoretically corresponding to trimeric (14 kDa), hexameric (27 kDa), nonameric (40 kDa) and dodecameric (56 kDa) $Aβ_{1-42}$ assemblies. These species represent multiples of trimeric Aβ oligomers, with high-molecular-mass assemblies (>20 kDa) appearing in mice older than 6 months. The detection of similar bands using 6E10 and 4G8 antibodies excludes the possibility that they represent degradation products of soluble APP, which lacks the mid-domain Aβ epitope ($Aβ_{17-24}$) recognized by 4G8 (Supplementary Fig. 2a). The bands were not recognized by 22C11 or APPC17-Cter antibodies, indicating they were neither APP nor APP cleavage-end products (data not shown).

Although this result suggests that ageing induces Aβ trimers to associate and form high-molecular-mass assemblies, we also considered the possibility that they might represent Aβ oligomers

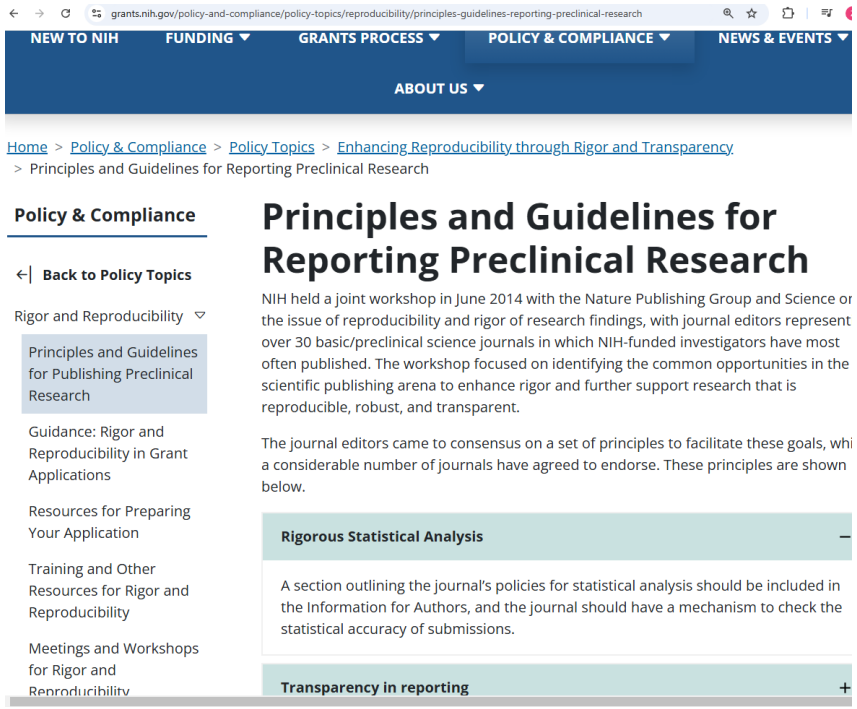
Nature **440**, 352–357 (2006) | Cite this article

Factors Contributing to Reproducibility Crisis

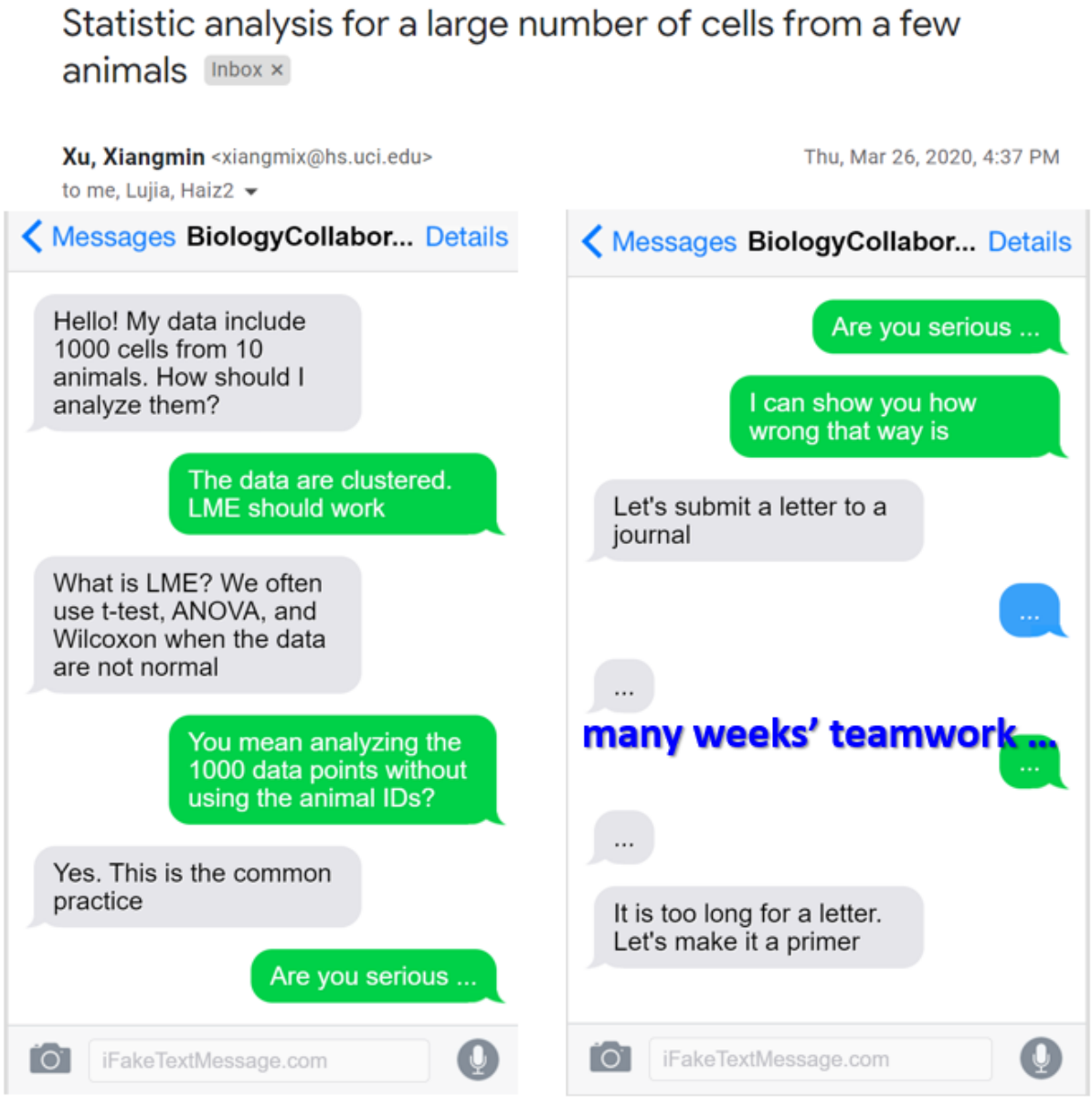
- Publication bias
- Hyper-competitive research environment
- Lack of adherence to good scientific practices
- Bad statistical practice

Factors Contributing to Reproducibility Crisis

- Bad statistical practice
  - underdeveloped experimental designs
  - over-interpretation of statistically marginal differences
  - lack of transparent presentation of methods
  - misuse of statistical methods



Awareness


Statistic analysis for a large number of cells from a few animals
Inbox ×
Xu, Xiangmin <xiangmix@hs.uci.edu>
Thu, Mar 26, 2020, 4:37 PM
to me, Lujia, Haiz2
Messages BiologyCollabor... Details
Hello! My data include 1000 cells from 10 animals. How should I analyze them?
The data are clustered. LME should work
What is LME? We often use t-test, ANOVA, and Wilcoxon when the data are not normal
You mean analyzing the 1000 data points without using the animal IDs?
Yes. This is the common practice
Are you serious ...
iFakeTextMessage.com
Messages BiologyCollabor... Details
Are you serious ...
I can show you how wrong that way is
Let's submit a letter to a journal
...
...
many weeks' teamwork ...
...
...
It is too long for a letter. Let's make it a primer
iFakeTextMessage.com

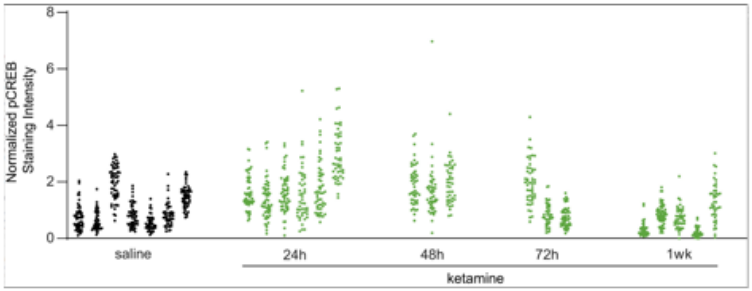
Better Statistical Practice?

- How is the situation compared to ten years ago?
- Aarts et al. (2014 Nature Neuroscience) examined 314 articles published Jan 2012 – June 2013 in five top journals: Science, Nature, Cell, Nature Neuroscience, Neuron: 53% involved nested (clustered) data but "conventional" analysis methods were used in all the articles.
- We surveyed articles published in prestigious journals over the past few years. Among >100 articles in which multiple observations were measured from each individual animal, <50% accounted for data dependencies in any meaningful way.

Examples of Wrong Analyses

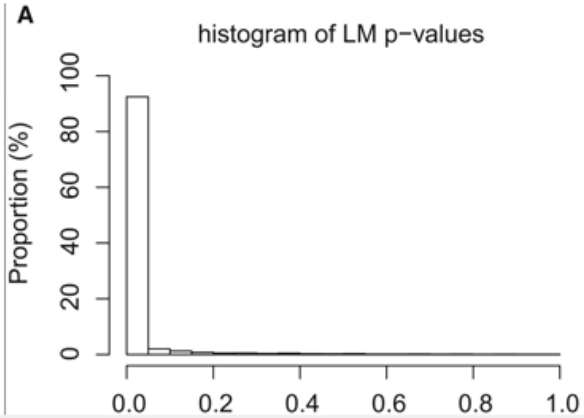
- "t(28656) = 314 with p<10^{-10} over a total of n=28657 neurons pooled across six mice"
- "n = 377 neurons from four mice, two-sided Wilcoxon signed rank test"
- "610 A cells, 987 B cells and 2584 C cells from 10 mice, one-way ANOVA and Kruskal–Wallis test"
- "two-sided paired t test, n=1597 neurons from 11 animals, d.f. = 1596"

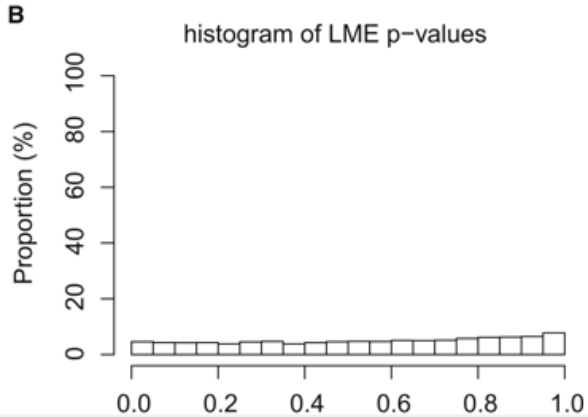
Attention Please!



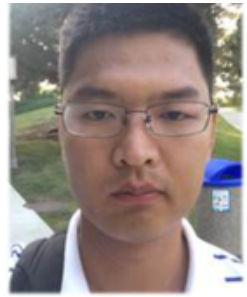
1200 neurons/measurements from 24 mice

	Saline (7 mice)	24 h (6 mice)	48 h (3 mice)	72 h (3 mice)	1 week (5 mice)
No. cells	357	209	139	150	245
ICC	0.61	0.33	0.02	0.63	0.54





Our First Action



Lujia Chen
PhD Candidate

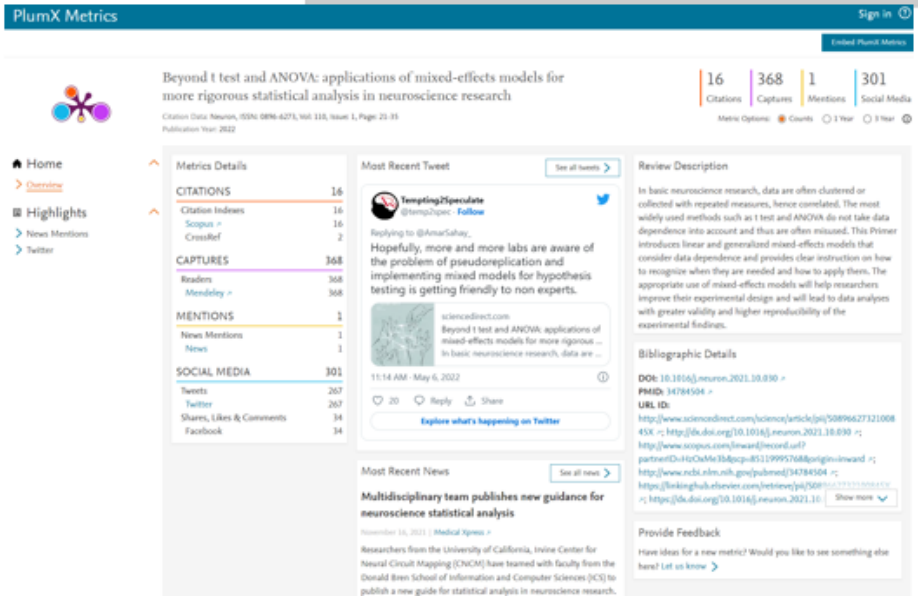
Steven Grieco

Postdoctoral Scholar

UCI School of Medicine

UCI's multidisciplinary team of researchers included (from left) Xiangmin Xu, PhD, professor and Chancellor's Fellow in the Department of Anatomy and Neurobiology, and director for the Center for Neural Circuit Mapping at the UCI School of Medicine; Zhaoxia Yu, PhD, and Michele Guindani, PhD, professors from the Department of Statistics UCI Donald Bren School of Information & Computer Sciences; and Todd Holmes, PhD, professor and vice chair of the UCI School of Medicine Department of Physiology & Biophysics.

James (Jim) Dent, LSSBB, DTM on LinkedIn

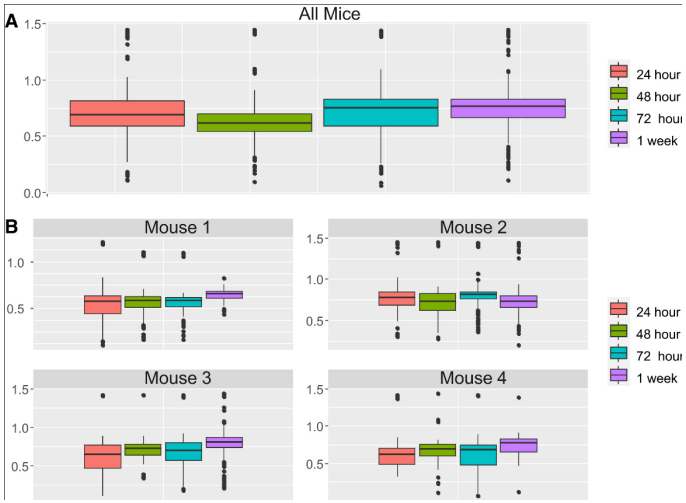


Accessible Materials

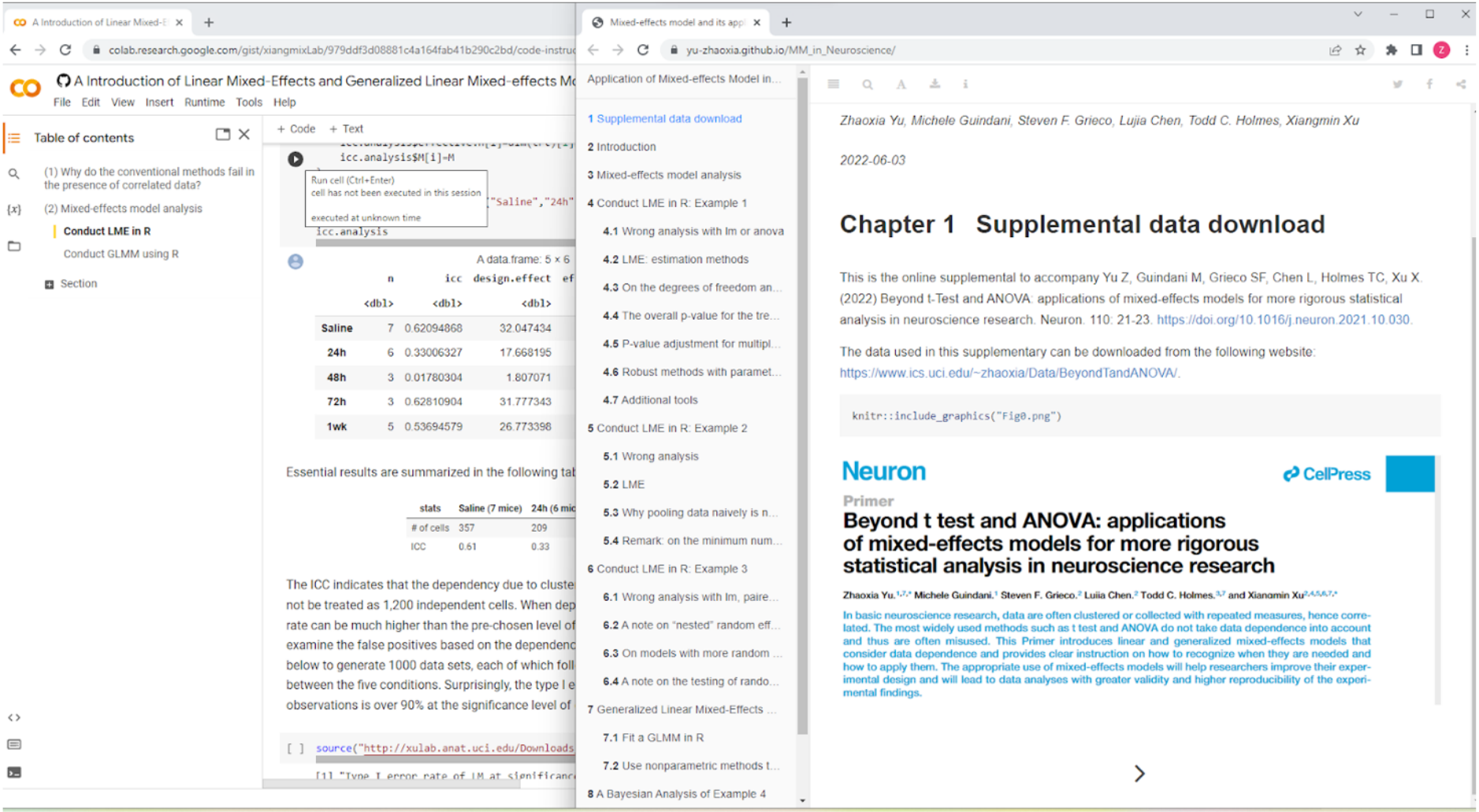
- Start from the methods that everyone understands
- T-test $\Rightarrow$ ANOVA $\Rightarrow$ LM $\Rightarrow$ LME $\Rightarrow$ GLMM $\Rightarrow$ Bayesian
- Representative examples and data
- Detailed supplemental materials
- Step-by-step guidance and reproducible code (including fixed random seeds)

Representative Examples

- Example 1: unacceptable type I error rate due to ignoring data dependency.
- Example 2: when cells are pooled naively, the results can be dominated by the animal contributing to a large proportion of cells.
- Example 3: nested random effects. P-values became more significant after accounting for data dependency.
- Example 4: non-normal outcomes.



Accessible Resources



Deliver Messages

- "Rigorous statistical analysis is not a hunt for the smallest p value (commonly known as p-hacking or significance chasing)."
- Overuse of p-values: "It should be noted that the modeling decisions should not be based on tests and p values alone, as the result may be significant even . . . or be insignificant . . . . Rather, the modeling decision should always be guided by the combined information provided by the study design, scientific reasoning, and previous evidence."
- "Due to the limited space, it is overambitious to cover all of the practical issues ... We refer the interested reader to specialized research articles (. . .) or to consult with experienced statisticians."

The Misuse of A Robust Test

Confronting false discoveries in single-cell differential expression

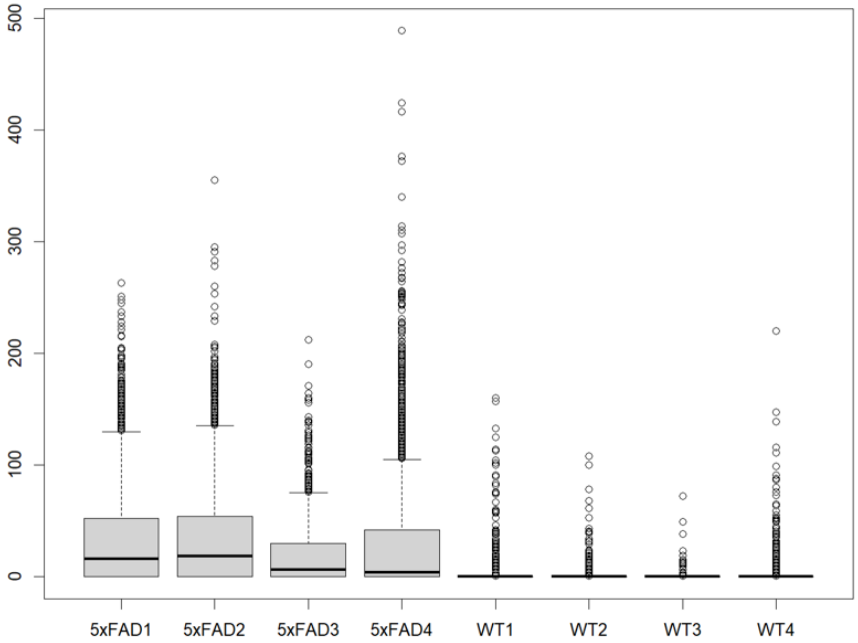
Jordan W. Squair, Matthieu Gautier, Claudia Kathe, Mark A. Anderson, Nicholas D. James, Thomas H. Hutson, Rémi Hudelle, Taha Qaiser, Kaya J. E. Matson, Quentin Barraud, Ariel J. Levine, Gioele La Manno, Michael A. Skinnider & Grégoire Courtine

Nature Communications **12**, Article number: 5692 (2021) | Cite this article

127k Accesses | **197** Altmetric | Metrics

endothelial cells using representative single-cell and pseudobulk methods. We selected the Wilcoxon rank-sum test as a single-cell method, since this test has been the most widely used approach in the field of single-cell transcriptomics (Fig. 1b), and edgeR-LRT[29] as a pseudobulk

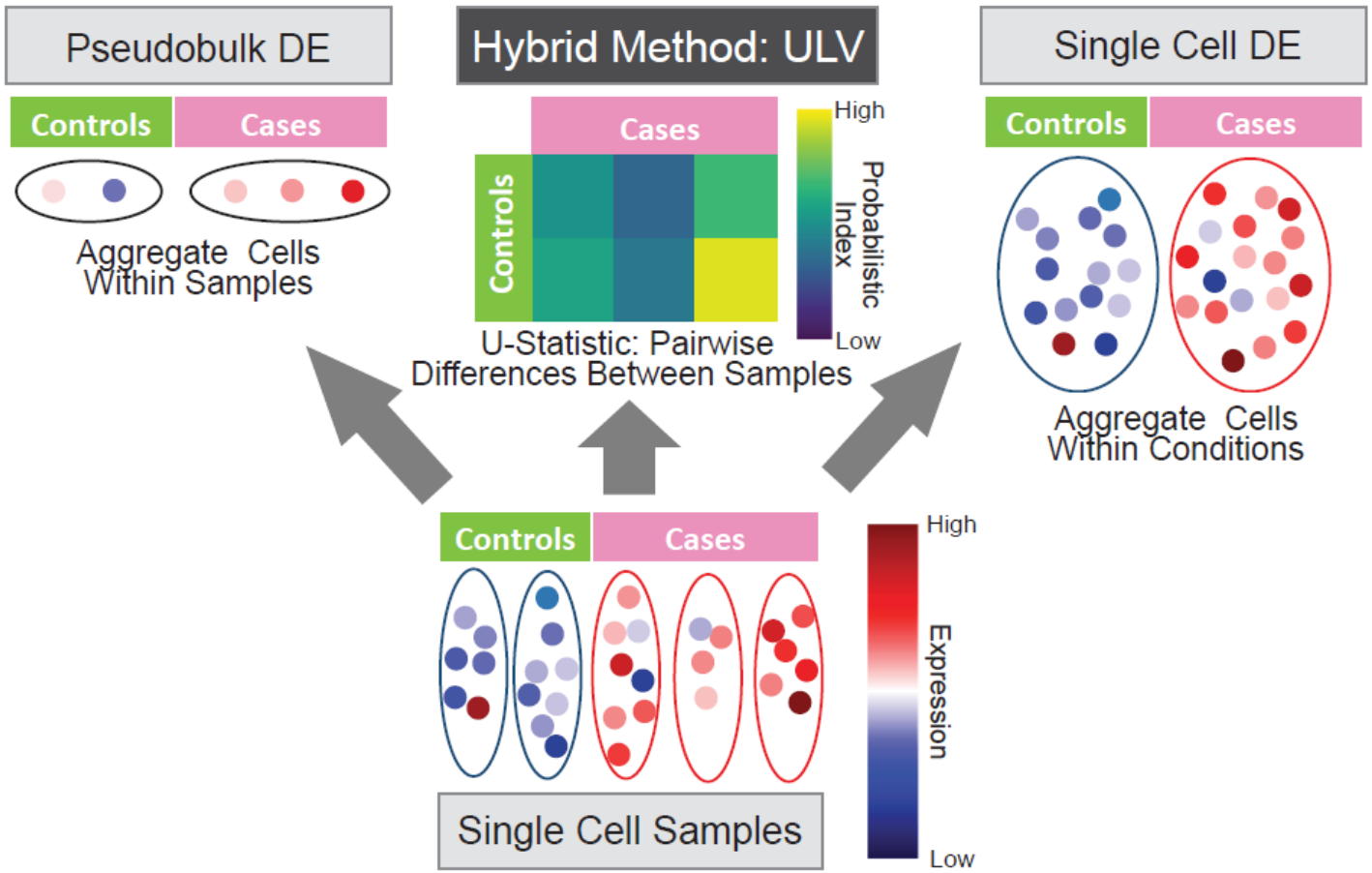
Why Nonprametric Tests are Appealing?



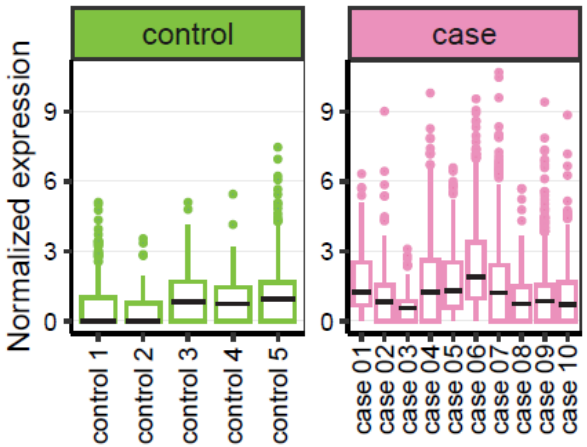
- Non-normal distributions, outliers, excess zeros
- Challenges: small sample sizes, large repeated measurements

Existing and Proposed Approaches

https://github.com/yu-zhaoxia/ULV



Goal: Robust, Valid, and Flexible



Subject Level Covariate Information

Subject ID	Age	Gender	...
case 1	63	Male	...
case 2	80	Female	...
...	...		
control 1	31	Female	...
control 2	45	Male	...
...	...	...	...

#case m = 10

#control n = 5

- Covariates

Existing Approaches

- Parametric methods: MAST-GLMM with zero-inflated Poisson or negative binomial, pseudobulk DeSeq2.
- Wilcoxon rank-sum test for cell data is the most widely used (Squair et al. 2021). But ...
- Wilcoxon rank-sum test for clustered data
  - use cluster-level summaries: when m=n=4, the smallest p-value is $1/70 \approx 0.014$
  - sample one observation per cluster at a time (Datta–Satten), large-sample approximations (Rosner et al.)
    - Large-sample sizes are required.
    - The number of repeated measurements per cluster is often small.
    - Better suited for longitudinal studies.
    - Not easy to extend to multiple groups and adjust for covariates

U Statistic and Rank Sum for Two Samples

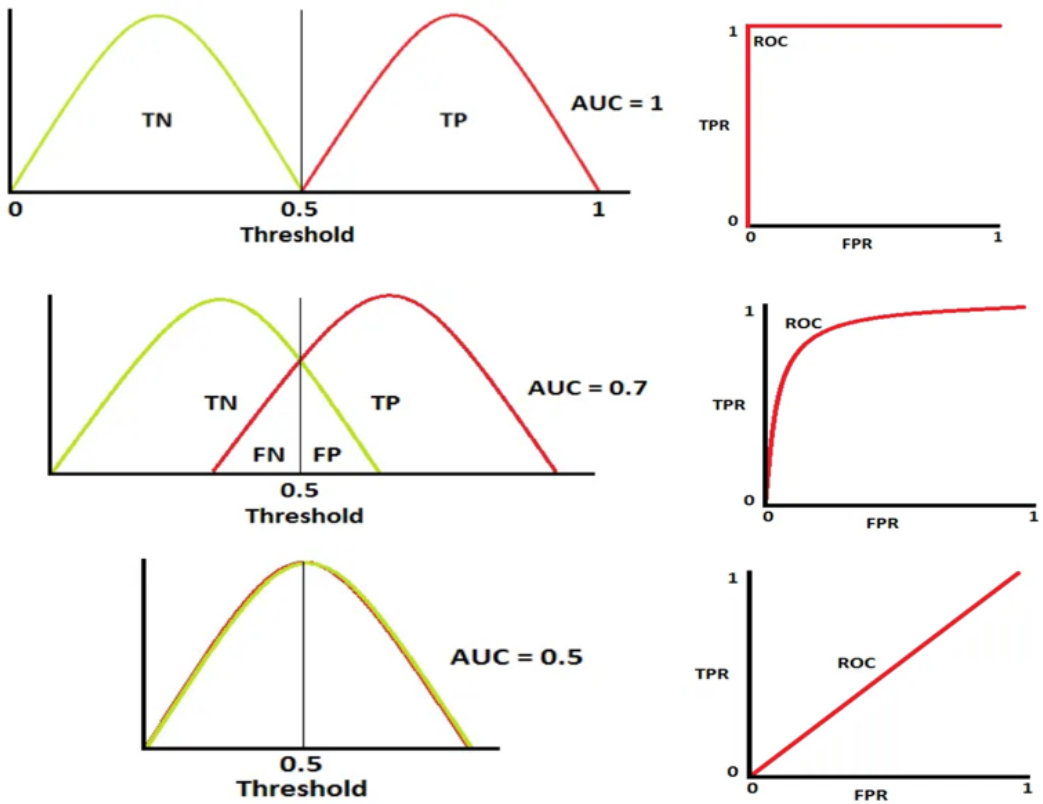
- Consider $X_1, \cdots, X_m$ and $Y_1, \cdots, Y_n$.
- Mann-Whitney U test is based on

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathrm{I}(X_i > Y_j)$$

- Wilcoxon rank sum test is based on R_X, the sum of ranks in group X.
- Mann-Whitney U and Wilcoxon rank sum are equivalent because

$$U = R_X - \frac{m(m+1)}{2}$$

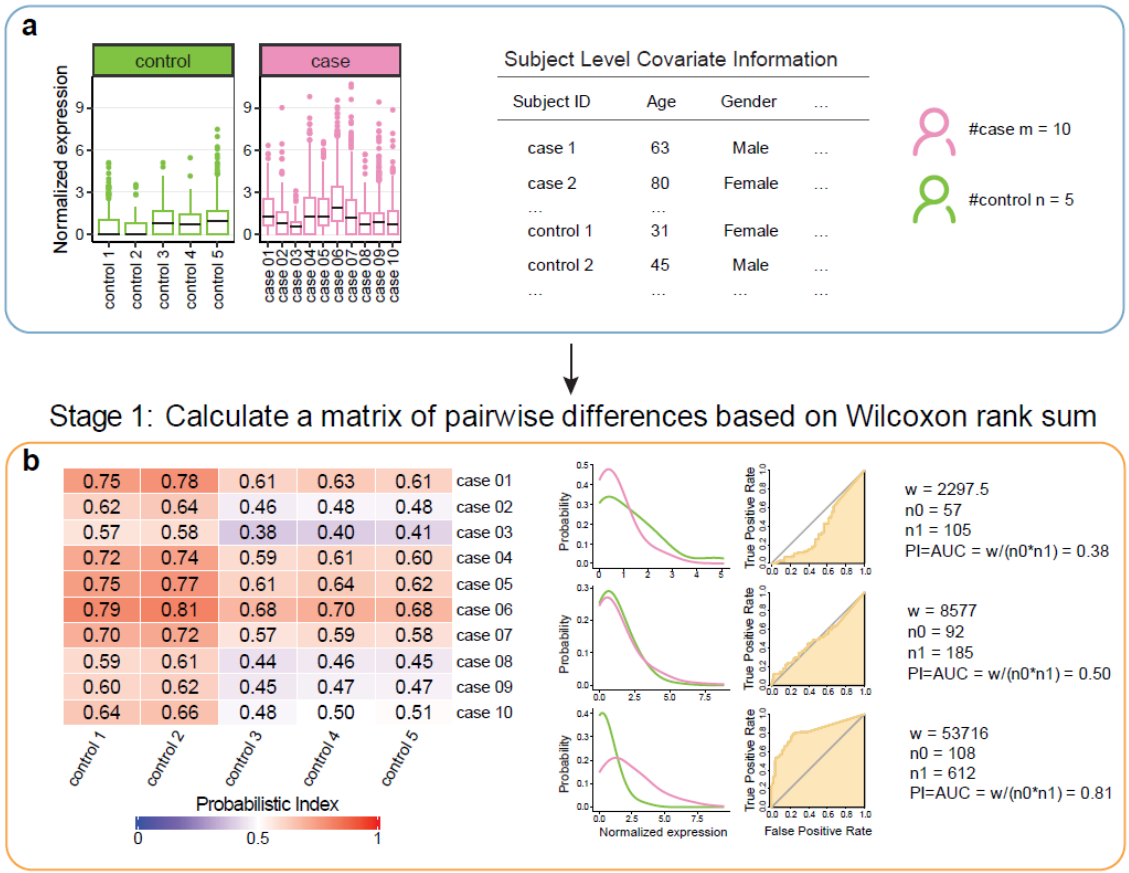
AUC as A Metric for Prediction / Separation



- $AUC = \frac{U}{mn}$

Quantify the Difference between Two Subjects

- AUC is also known as the **Probabilistic Index (PI)**



Two-Group Comparisons: Notations

- Consider m cases and n controls, each with multiple measurements.
- The ith case is denoted by

$$\vec{y}_{1i} = (y_{1,i,1}, \cdots, y_{1,i,K_{1i}})^T,$$

  where K_{1i} is the number repeated measurements.
- Similarly, the jth control is denoted by

$$\vec{y}_{0j} = (y_{0,j,1}, \cdots, y_{0,j,K_{0j}})^T$$

- The comparison between the ith case and jth control is given by $h(\vec{y}_{1i}, \vec{y}_{0j})$, which is a U statistic.

Two-Group Comparisons

- For example, one robust choice of h is

$$h(\vec{y}_{1i}, \vec{y}_{0j}) = \frac{1}{K_{1i}K_{0j}} \sum_{k=1}^{K_{1i}} \sum_{k'=1}^{K_{0j}} \left[I(y_{1ik} > y_{0jk'}) + \frac{1}{2} I(y_{1ik} = y_{0jk'}) \right].$$

- It is a consistent, unbiased, and minimum variance estimator of the probabilistic index for quantifying "greater".
- A reasonable statistic is $U_{cluster} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{=1}^{n} h(\vec{y}_{1i}, \vec{y}_{0j})$
- $U_{cluster}$ can be used for inference. However, its large-sample distribution requires estimating correlations, so sample sizes need to be large.

A Latent Variable Model

- Let $d_{ij} = h(\vec{y}_{1i}, \vec{y}_{0j}), i = 1, \cdots, m; j = 1, \cdots, n$
- We model d_{ij} by

$$d_{ij} = a_i - b_j + \epsilon_{ij}, i = 1, \cdots, m; j = 1, \cdots, n$$
$$a_i \sim N(\mu, \sigma_1^2), b_j \sim N(0, \sigma_0^2), \epsilon_{ij} \sim N(0, \sigma^2)$$

- a_i represents the latent (relative and not directly observed) level of the ith case
- b_j is the latent level of the jth control
- μ is the difference in the population levels of the two groups, which is the parameter of interest.

$$H_0 : \mu = 0.5$$

Justification of the Latent Variable Model

- The model includes the two-sample t-test as a special case when $d(i,j) = \bar{x}_i - \bar{y}_j$.
- Connection to Rasch (1 parameter Item Response Model) Model
  - Proposed by Rasch in 1960 for modeling responses of test or survey items.
  - The Rasch model and extensions are widely used in
    - psychometrics (van der Linden & Hamble- ton, 2013)
    - educational tests (Lan et al., 2016)
    - public health (Cappelleri et al., 2014)
    - market and financial research (Schellhorn & Sharma, 2013; Brzezi'nska, 2016).

Justification of the Latent Variable Model

- Let $Y_{ij} \in \{0, 1\}$ denote test-taker i's response to item j, with 1 for a correct response and 0 otherwise.
- The Rasch model assumes that

$$log\left(\frac{Pr(Y_{ih} = 1)}{Pr(Y_{ih} = 0)}\right) = \theta_i - \beta_j$$

- The original Rasch model assumes that the ability scores (θ_i) are random and the item difficulties (β_j) are fixed.

Extensions

- Adjust for covariates
- Compare multiple groups
- Weighted analysis to account for unbalanced cluster sizes

Adjust for Covariates

- Suppose there are p covariates for each subject, $x_i = (x_{i1}, \cdots x_{ip})^T$,

$$\begin{aligned} d_{ij} &= (a_{0i} + \beta^T x_i) - (b_{0j} + \beta^T x_j) + \epsilon_{ij} \\ &= a_{0i} - b_{0j} + \beta^T (x_i - x_j) + \epsilon_{ij}, \end{aligned}$$

- a_{0i} and b_{0j} represent the covariate-adjusted latent expression levels.

$$a_{0i} \sim N(\mu, \sigma_1^2), b_{0j} \sim N(0, \sigma_0^2), \epsilon_{ij} \sim N(0, \sigma^2)$$

- Testing $H_0 : \mu = 0.5$.

Compare Multiple Groups

- Suppose there are a total number of $M + 1$ conditions.
- The pairwise difference between the observations from the ith subject under condition m and the jth subject in the reference group is

$$d_{ij}^{m} = a_{mi} - b_{0j} + \epsilon_{ij}, i = 1, \cdots, N_m; j = 1, \cdots, N_0; m = 1, \cdots, M,$$

- a_{mi} represents the latent level of the ith subject in condition m, and b_{0j} is the latent level of the jth subject in the reference group.

$$a_{mi} \sim N(\mu_m, \sigma_m^2), b_{0j} \sim N(0, \sigma_0^2), \epsilon_{ij} \sim N(0, \sigma_\epsilon^2).$$

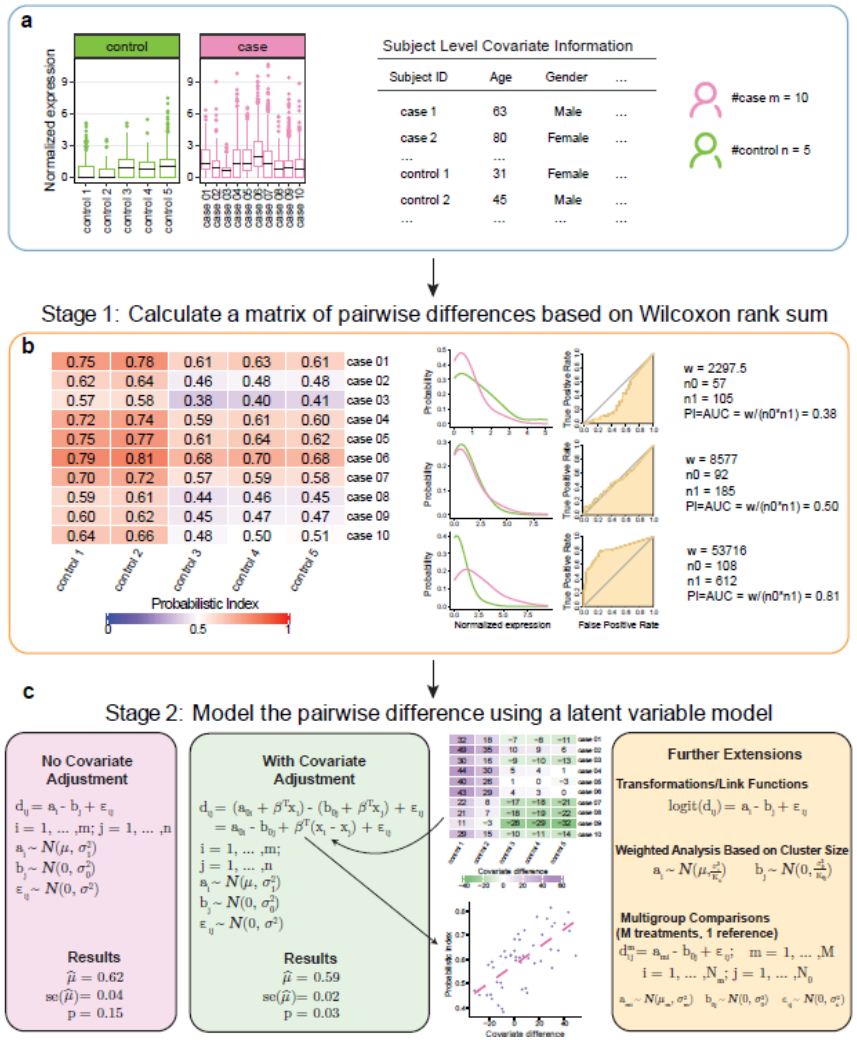
- Testing $H_0 : \mu_1 = \cdots = \mu_M = 0.5$.

Weighted Analysis

- Let K_{0i} and K_{1j} denote the cluster sizes of control i and patient j, respectively.
- $d_{ij} \propto (1/K_{0i} + 1/K_{1j})$
- Weighted latent variable model:

$$a_i \sim N(\mu, \frac{\sigma_1^2}{K_{1i}}), b_j \sim N(0, \frac{\sigma_0^2}{K_{0j}}).$$

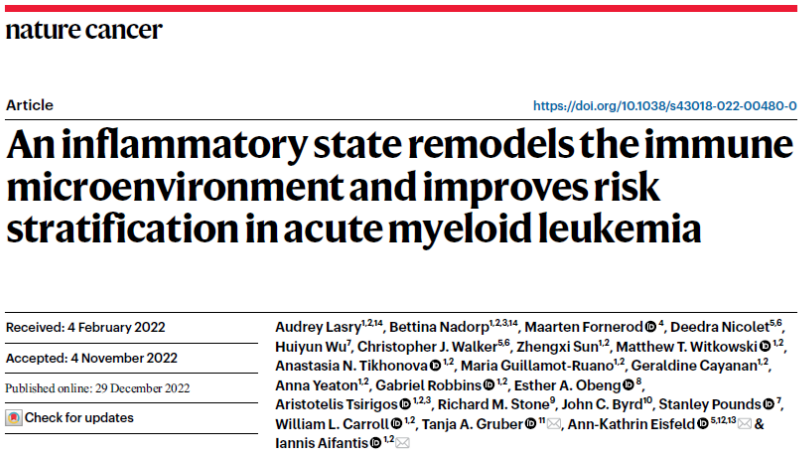
A Schematic Summary of ULV



Simulation Results

- We simulated single-cell expression data based on parameters estimated from real sc RNA-seq data
- ULV has excellent control of type I error rates at different significance levels
- ULV has comparable power with competing parametric or pseudobulk methods
- ULV is much more powerful than the pseudobulk version of Wilcoxon rank sum test

Application 1: ULV discovers DE proteins in AML patients



nature cancer

Article

https://doi.org/10.1038/s43018-022-00480-0

An inflammatory state remodels the immune microenvironment and improves risk stratification in acute myeloid leukemia

Received: 4 February 2022

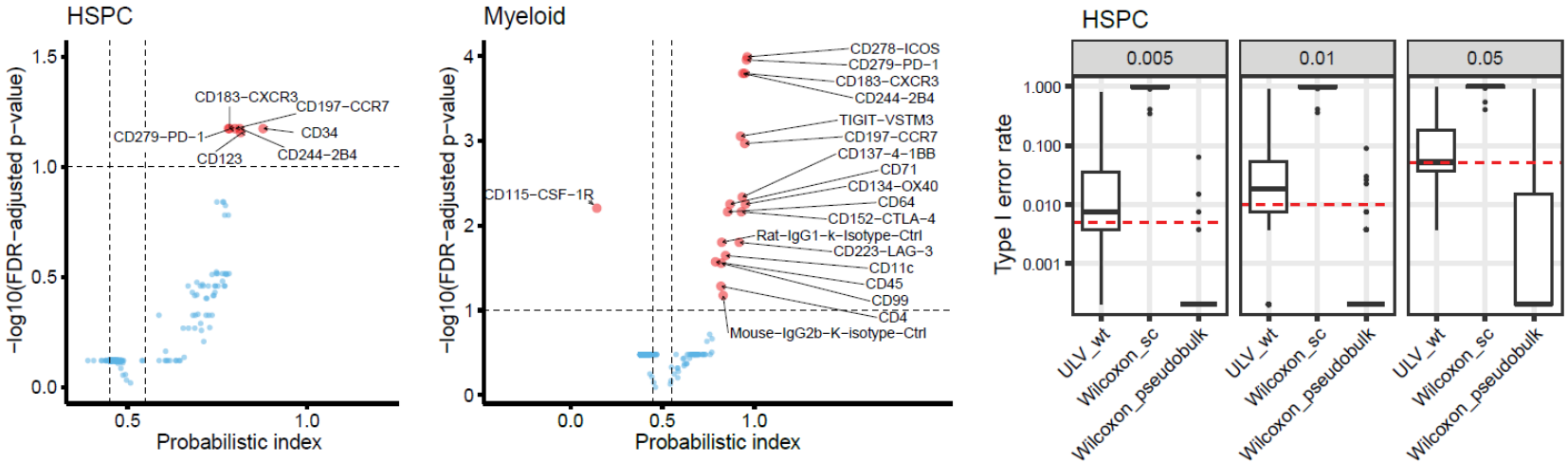
Accepted: 4 November 2022

Published online: 29 December 2022

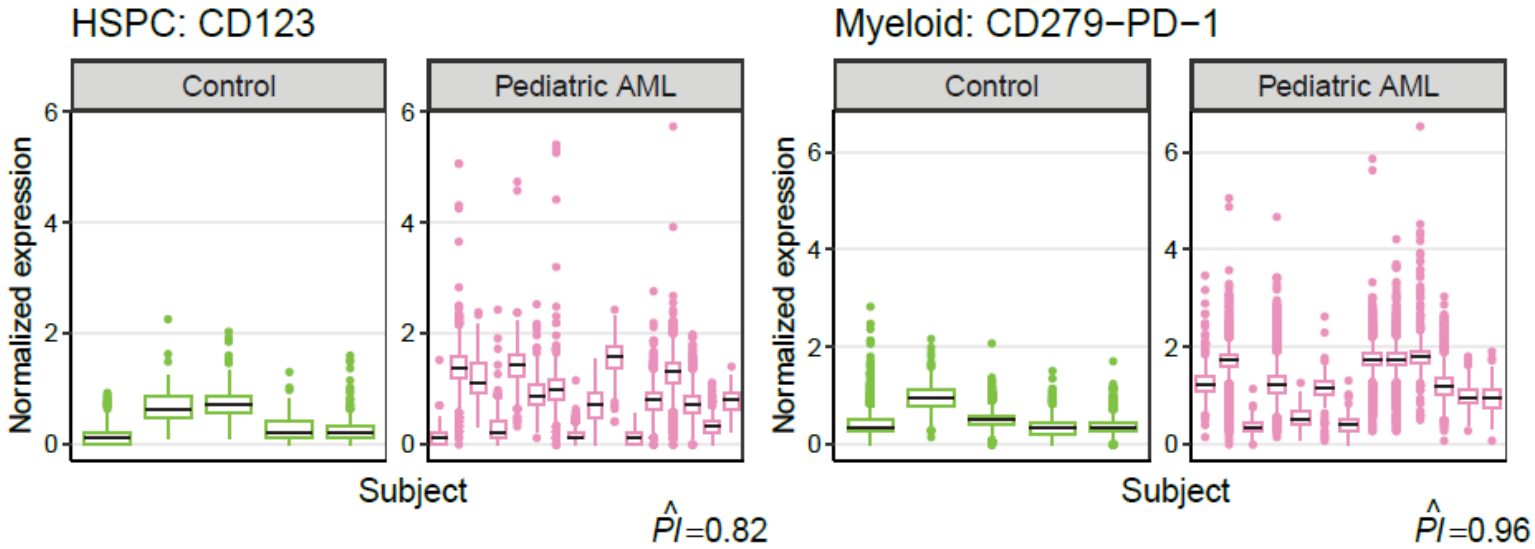
Check for updates

Audrey Lasry[1,2,14], Bettina Nadorp[1,2,3,14], Maarten Fornerod[4], Deedra Nicolet[5,6], Huiyun Wu[7], Christopher J. Walker[5,6], Zhengxi Sun[1,2], Matthew T. Witkowski[1,2], Anastasia N. Tikhonova[1,2], Maria Guillamot-Ruano[1,2], Geraldine Cayanan[1,2], Anna Yeaton[1,2], Gabriel Robbins[1,2], Esther A. Obeng[8], Aristotelis Tsirigos[1,2,3], Richard M. Stone[9], John C. Byrd[10], Stanley Pounds[7], William L. Carroll[1,2], Tanja A. Gruber[11] ✉, Ann-Kathrin Eisfeld[5,12,13] ✉ & Iannis Aifantis[1,2] ✉

- We focused on two cell types:
  - 56,764 malignant and healthy hematopoietic stem and progenitor cells (HSPCs): 269 proteins, 6 pediatric AML patients and 5 age-matched controls
  - 65,510 myeloid cells: 270 proteins, 13 pediatric AML patients and 5 age-matched controls
- Protein expression data were pre-processed using centered log-ratio normalization

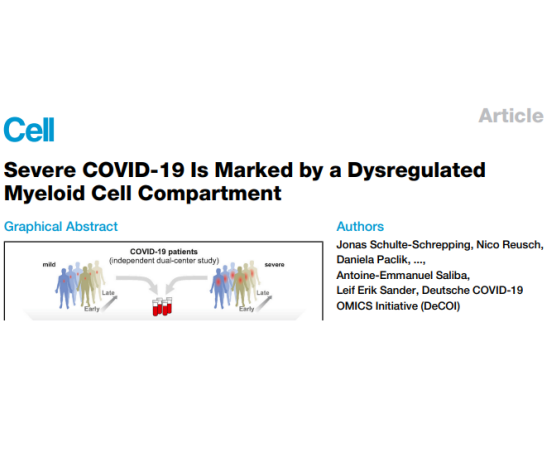


- Pseudobulk Wilcoxon didn't find any DE protein



- The interleukin 3 receptor alpha (CD123) was previously found as a putative marker in AML blasts compared with normal HSPCs
- The programmed cell death protein 1 (CD279-PD-1) has been linked to the suppression of immune responses during tumor development and is also associated with AML pathology

Application 2: COVID-19 Data


Cell

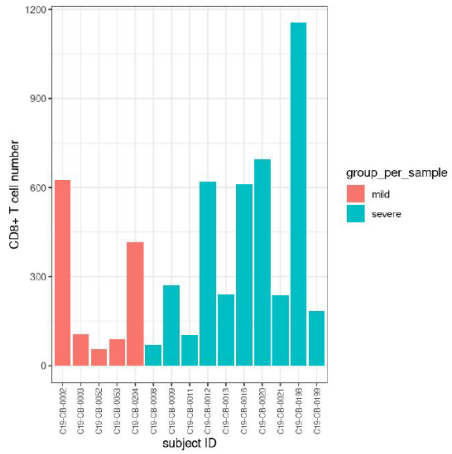
Article

Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment

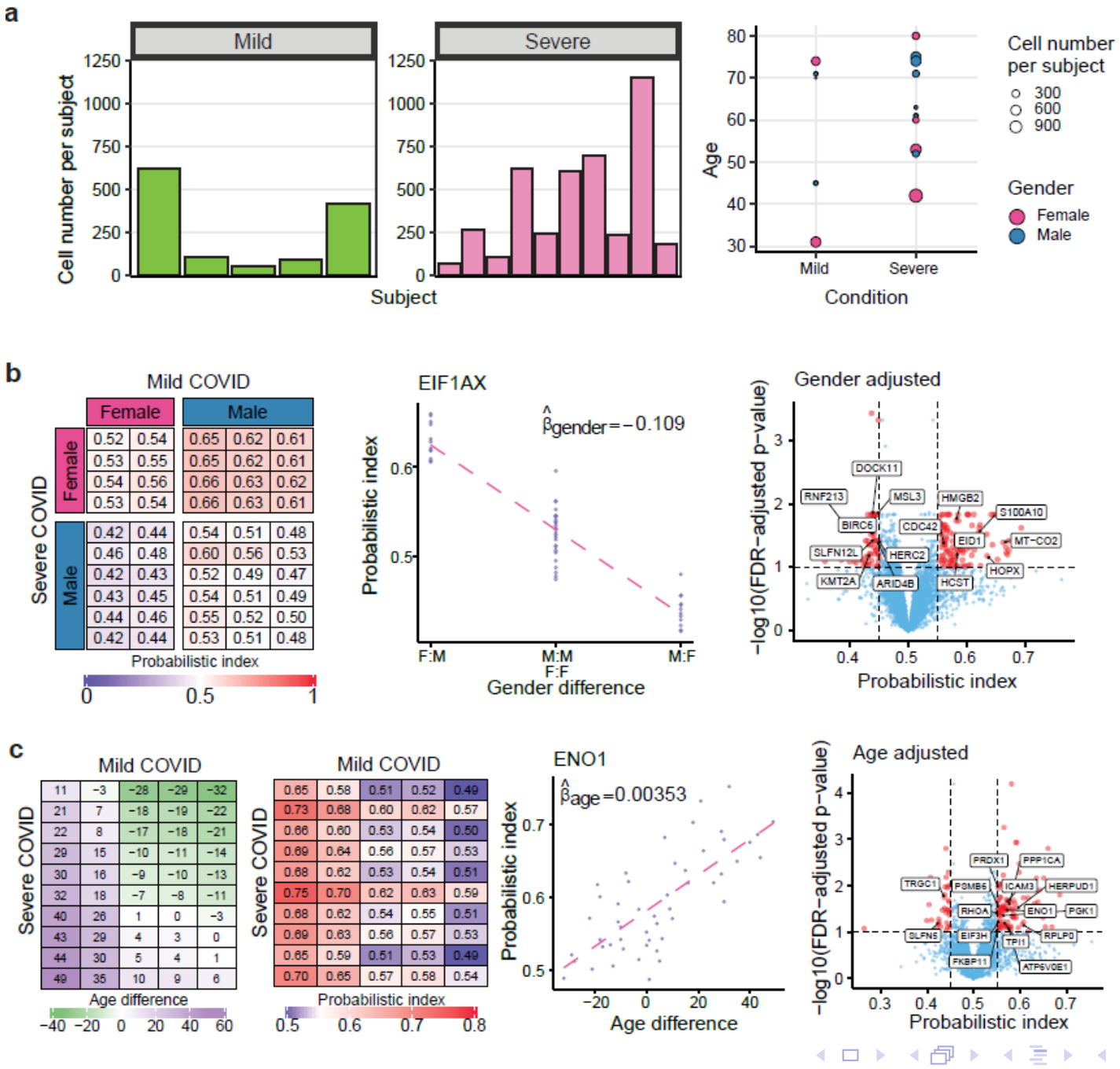
Graphical Abstract

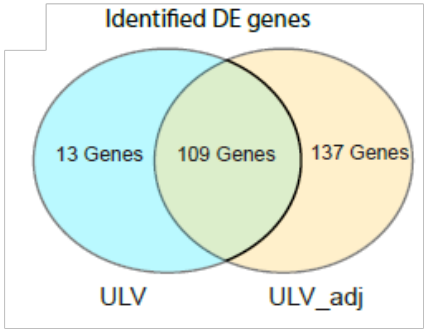
Authors

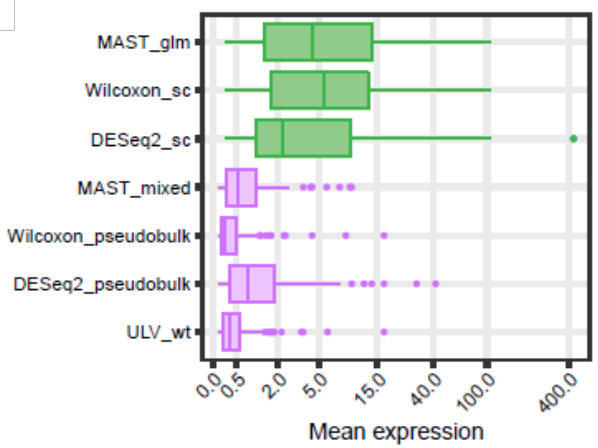
Jonas Schulte-Schrepping, Nico Reusch, Daniela Paclik, ..., Antoine-Emmanuel Saliba, Leif Erik Sander, Deutsche COVID-19 OMICS Initiative (DeCOI)

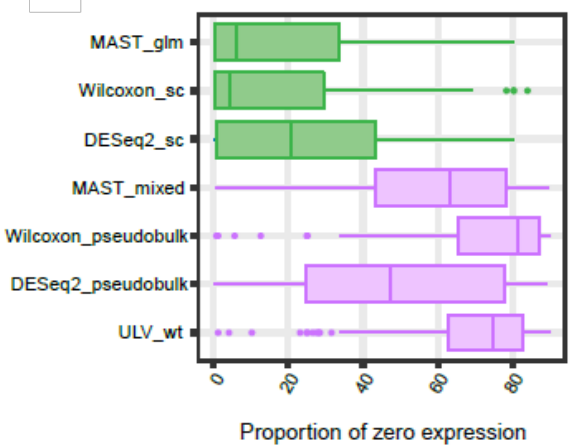


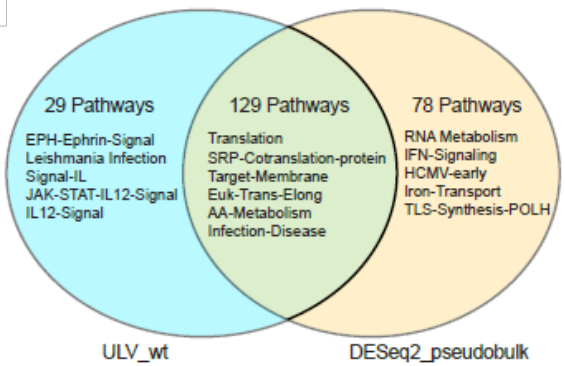
- Blood samples from Germany COVID-19 patients.
- scRNA-seq data were measured using a droplet-based single-cell platform (10x Chromium), which contains the peripheral blood mononuclear cells (PBMC).
- We focused on patients with an adequate number of CD8+ T cells: 5 mild and 10 severe patients.

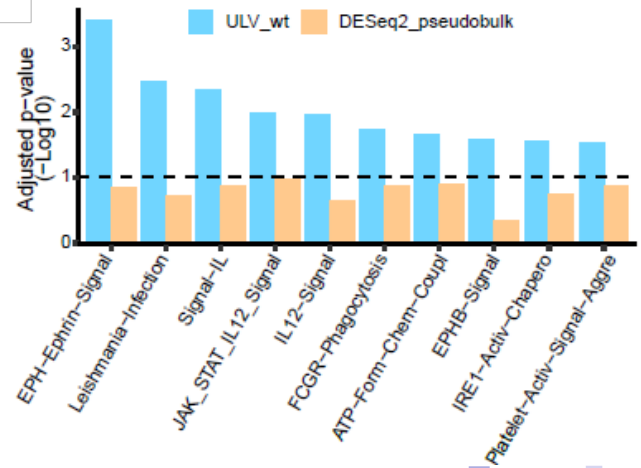

a
Mild
Severe
Cell number per subject
Subject
Age
Condition
Cell number per subject
300
600
900
Gender
Female
Male
b
Mild COVID
Female
Male
Severe COVID
Probabilistic index
EIF1AX
β̂gender=−0.109
Probabilistic index
F:M
M:M
F:F
M:F
Gender difference
Gender adjusted
−log10(FDR−adjusted p−value)
Probabilistic index
DOCK11
RNF213
MSL3
HMGB2
S100A10
BIRC6
CDC42
EID1
MT−CO2
SLFN12L
HERC2
HOPX
KMT2A
ARID4B
HCST
c
Mild COVID
Severe COVID
Age difference
Mild COVID
Severe COVID
Probabilistic index
ENO1
β̂age=0.00353
Probabilistic index
Age difference
Age adjusted
−log10(FDR−adjusted p−value)
Probabilistic index
PRDX1
PPP1CA
TRGC1
PSMB6
ICAM3
HERPUD1
RHOA
ENO1
PGK1
SLFN5
EIF3H
TPI1
RPLP0
FKBP11
ATP6V0E1


Identified DE genes
13 Genes
109 Genes
137 Genes
ULV
ULV_adj


MAST_glm
Wilcoxon_sc
DESeq2_sc
MAST_mixed
Wilcoxon_pseudobulk
DESeq2_pseudobulk
ULV_wt
0.0
0.5
2.0
5.0
15.0
40.0
100.0
400.0
Mean expression


MAST_glm
Wilcoxon_sc
DESeq2_sc
MAST_mixed
Wilcoxon_pseudobulk
DESeq2_pseudobulk
ULV_wt
0
20
40
60
80
Proportion of zero expression


29 Pathways
EPH-Ephrin-Signal
Leishmania Infection
Signal-IL
JAK-STAT-IL12-Signal
IL12-Signal
129 Pathways
Translation
SRP-Cotranslation-protein
Target-Membrane
Euk-Trans-Elong
AA-Metabolism
Infection-Disease
78 Pathways
RNA Metabolism
IFN-Signaling
HCMV-early
Iron-Transport
TLS-Synthesis-POLH
ULV_wt
DESeq2_pseudobulk

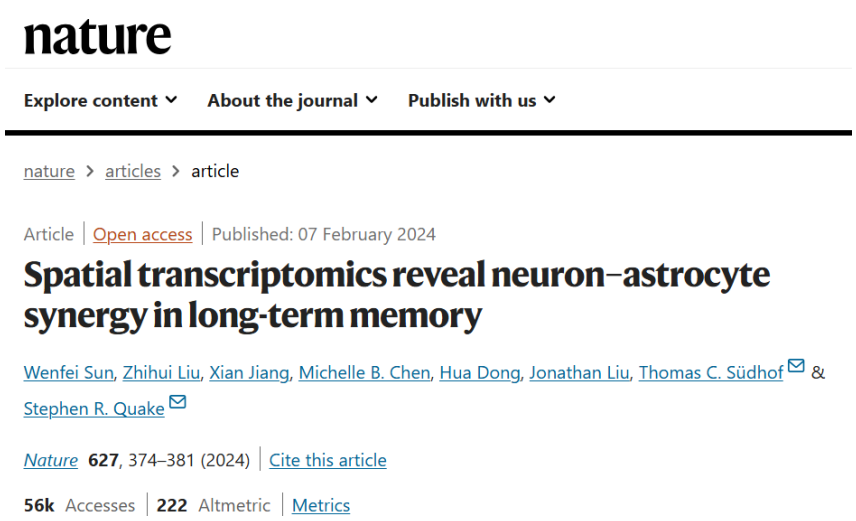

ULV_wt
DESeq2_pseudobulk
Adjusted p-value
(−Log10)
0
1
2
3
EPH−Ephrin−Signal
Leishmania−Infection
Signal−IL
JAK_STAT_IL12_Signal
IL12−Signal
FCGR−Phagocytosis
ATP−Form−Chem−Coupl
EPHB−Signal
IRE1−Activ−Chapero
Platelet−Activ−Signal−Aggre

Misuse of Statistical Methods

- It is easy and sometimes attempting to misuse methods
- Can happen in many journals, including the most prestigious journals
- High impact journals have an enormous amount readers
- High impact journals should be more responsible to set good examples

Misuse of Statistical Methods

- Used the conventional rank sum test as a “robust" method
- No control of multiple comparisons:
  - reported 107 out 3350 genes were “significant" (nominal $p<0.05$, fold change >1.75)
  - But $5\% \times 3350 = 167$



nature

Explore content | About the journal | Publish with us

nature › articles › article

Article | Open access | Published: 07 February 2024

Spatial transcriptomics reveal neuron–astrocyte synergy in long-term memory

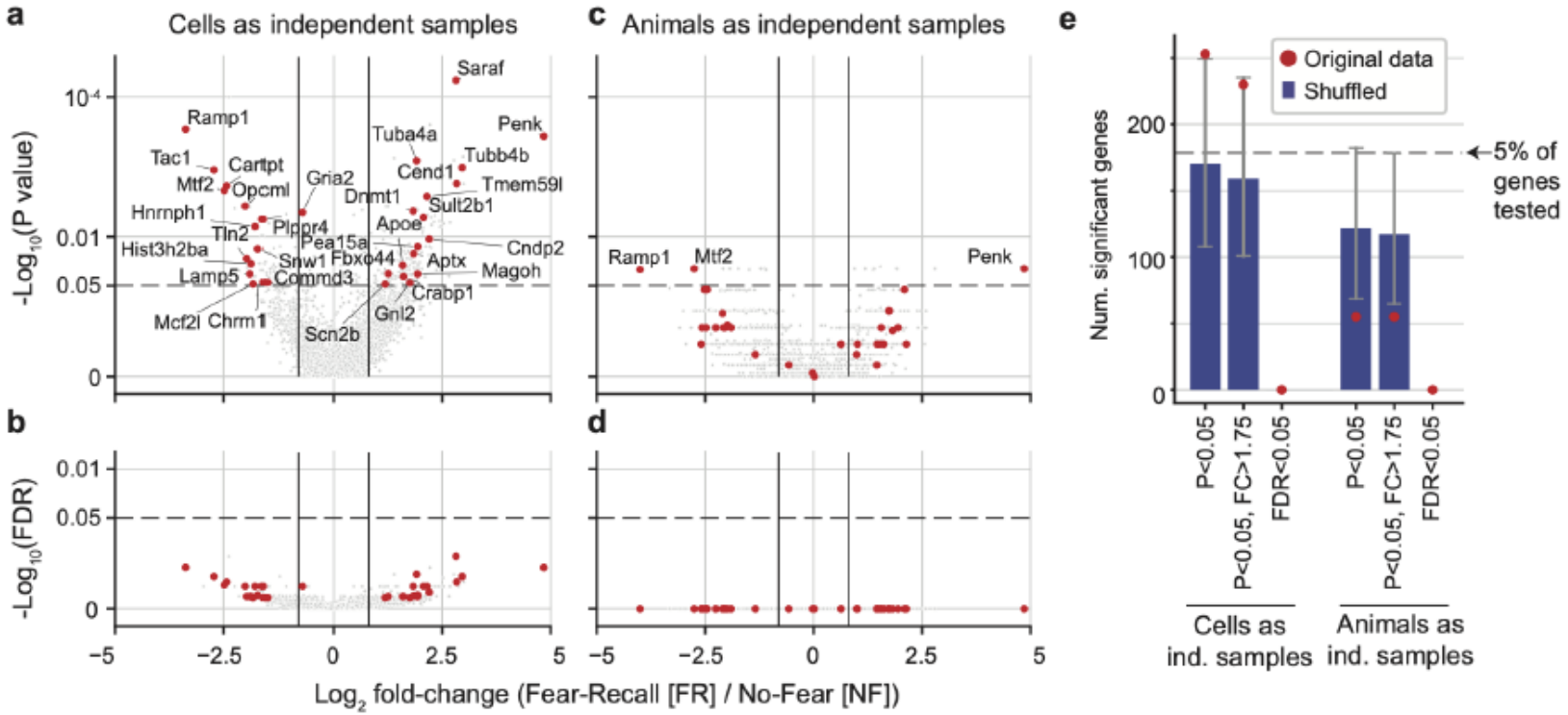
Wenfei Sun, Zhihui Liu, Xian Jiang, Michelle B. Chen, Hua Dong, Jonathan Liu, Thomas C. Südhof & Stephen R. Quake

Nature **627**, 374–381 (2024) | Cite this article

56k Accesses | **222** Altmetric | Metrics

Misuse of Statistical Methods

- Failing to adjust for multiple comparisons or correlated data leads to spurious DE genes



Discussion

- Misuse of statistical methods has contributed to the reproducibility crisis.
- Raise awareness, take action, and continue advocacy efforts.
- Collaborate with domain scientists from diverse backgrounds.
- Provide accessible materials and resources
- Understand the challenges and develop new methods and tools

Acknowledgements

- Student: Mingyu Du
- Collaborators:
  - Veronica Berrocal
  - Michele Guindani
  - Kevin Johnston (Utah Tech University), Steven F Grieco, Todd C Holmes, Wei Li, and Xingmin Xu, Eran Mukamel (UCSD)

Appendices

- Appendix 1: Justification of the ULV
- Appendix 2: A cautionary note
- Appendix 3: Simulation methods and results

Appendix 1: Justification

- As a two-way random-effects model
- Closed-form
- Least square

As a Two-Way Random-Effects Model

- This is also a two-way random-effects model with a single observation per cell.
- $\bar{d}_{..}$ is the LME and is unbiased for μ.
- The variance of $\hat{\mu} = \bar{d}_{..}$ can be estimated using

$$\hat{var}(\hat{\mu}) = \frac{1}{m}\hat{\sigma}_1^2 + \frac{1}{n}\hat{\sigma}_0^2 + \frac{1}{mn}\hat{\sigma}^2,$$

where

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{m}(\bar{d}_{i.} - \bar{d}_{..})^2}{m-1}, \hat{\sigma}_0^2 = \frac{\sum_{j=1}^{n}(\bar{d}_{.j} - \bar{d}_{..})^2}{n-1},$$

$$\hat{\sigma}^2 = \frac{1}{(m-1)(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}(d_{ij} - \bar{d}_{i.} - \bar{d}_{.j} + \bar{d}_{..})^2.$$

- The third item is small and can be ignored.

Closed-form

Theorem: Let $\bar{d}_{i\cdot}, i = 1, \cdots, m$ and $\bar{d}_{\cdot j}, j = 1, \cdots, n$ be the marginal means. Then

- $\bar{d}_{i\cdot}, i = 1, \cdots, m$ are identically distributed; so are $\bar{d}_{\cdot j}, j = 1, \cdots, n$.
- All the above means are asymptotically uncorrelated.
- Under the null hypothesis of no group difference

$$\frac{\bar{d}_{\cdot\cdot} - 0.5}{\sqrt{\hat{var}(\bar{d}_{\cdot\cdot})}} \to N(0, 1)$$

where

$$\hat{var}(\bar{d}_{\cdot\cdot}) = \frac{1}{m}\hat{\sigma}_1^2 + \frac{1}{n}\hat{\sigma}_0^2, \hat{\sigma}_1^2 = \frac{\sum_{i=1}^{m}(\bar{d}_{i\cdot} - \bar{d}_{\cdot\cdot})^2}{m-1}, \hat{\sigma}_0^2 = \frac{\sum_{j=1}^{n}(\bar{d}_{\cdot j} - \bar{d}_{\cdot\cdot})^2}{n-1}$$

Least Squares

- **Remark**: the above t-statistic is identical to the two-sample t-statistic based on $(\hat{a}_1, \cdots, \hat{a}_m)$ and $(\hat{b}_1, \cdots, \hat{b}_n)$, where

$$\hat{a}_i = \bar{d}_{i\cdot}, \hat{b}_j = \bar{d}_{\cdot j} - \bar{d}_{\cdot\cdot}, i = 1, \cdots, m; j = 1, \cdots, n,$$

  which is the least squares solution to
  $d_{ij} = a_i - b_j + \epsilon_{ij}, i = 1, \cdots, m; j = 1, \cdots, n$ under the constraint that $\sum_{j=1}^{b} b_j = 0$.
- Least squares solutions exist and are unique up to a constant.
- These closed forms can be used when there is no need to adjust for covariates.

Appendix 2: A Note

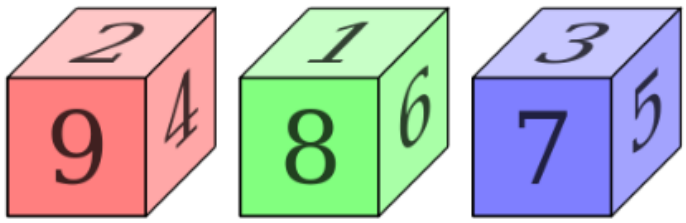
- Suppose $Pr(G_2 > G_0) > \frac{1}{2}$ and $Pr(G_0 > G_1) > \frac{1}{2}$
- What can we say about $Pr(G_2 > G_1)$?

A Cautionary Note About Rank-based Methods for Multiple Groups

- It turns out that

$$\left.\begin{array}{l} Pr(G_2 > G_0) > \frac{1}{2} \\ Pr(G_0 > G_1) > \frac{1}{2} \end{array}\right\} \not\Longrightarrow Pr(G_2 > G_1) > \frac{1}{2}$$

- Nontransitive dice



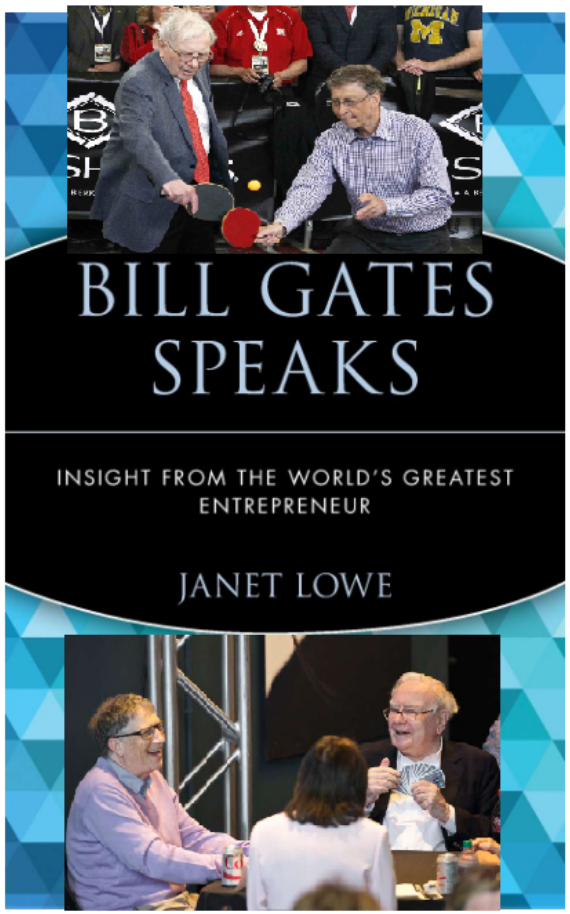
An example of intransitive dice (opposite sides have the same value as those shown).

- A: 2, 2, 6, 6, 7, 7
- B: 1, 1, 5, 5, 9, 9
- C: 3, 3, 4, 4, 8, 8

$P(A > B) = P(B > C) = P(C > A) = \frac{5}{9}$

A Cautionary Note About Rank-based Methods for Multiple Groups

- A story between Warren Buffet and Bill Gates



Buffett once challenged Gates to a game of dice, using a set of four unusual dice with a combination of numbers from 0 to 12 on the sides. Buffett suggested that each of them choose one of the dice, then discard the other two. They would bet on who would roll the highest number most often. Buffett offered to let Gates pick his die first. This suggestion instantly aroused Gates's curiosity. He asked to examine the dice, after which he demanded that Buffett choose first.

"It wasn't immediately evident that because of the clever selection of numbers for the dice they were nontransitive," Gates said. "The mathematical principle of transitivity, that if A beats B and B beats C, then A beats C, did not apply. Assuming ties were

A Cautionary Note About Rank-based Methods for Multiple Groups

- Nontransitivity could cause inconsistencies when comparing multiple conditions / treatments.
- The inference about overall effects is still valid.
- When two specific groups need to be compared, compare them directly.

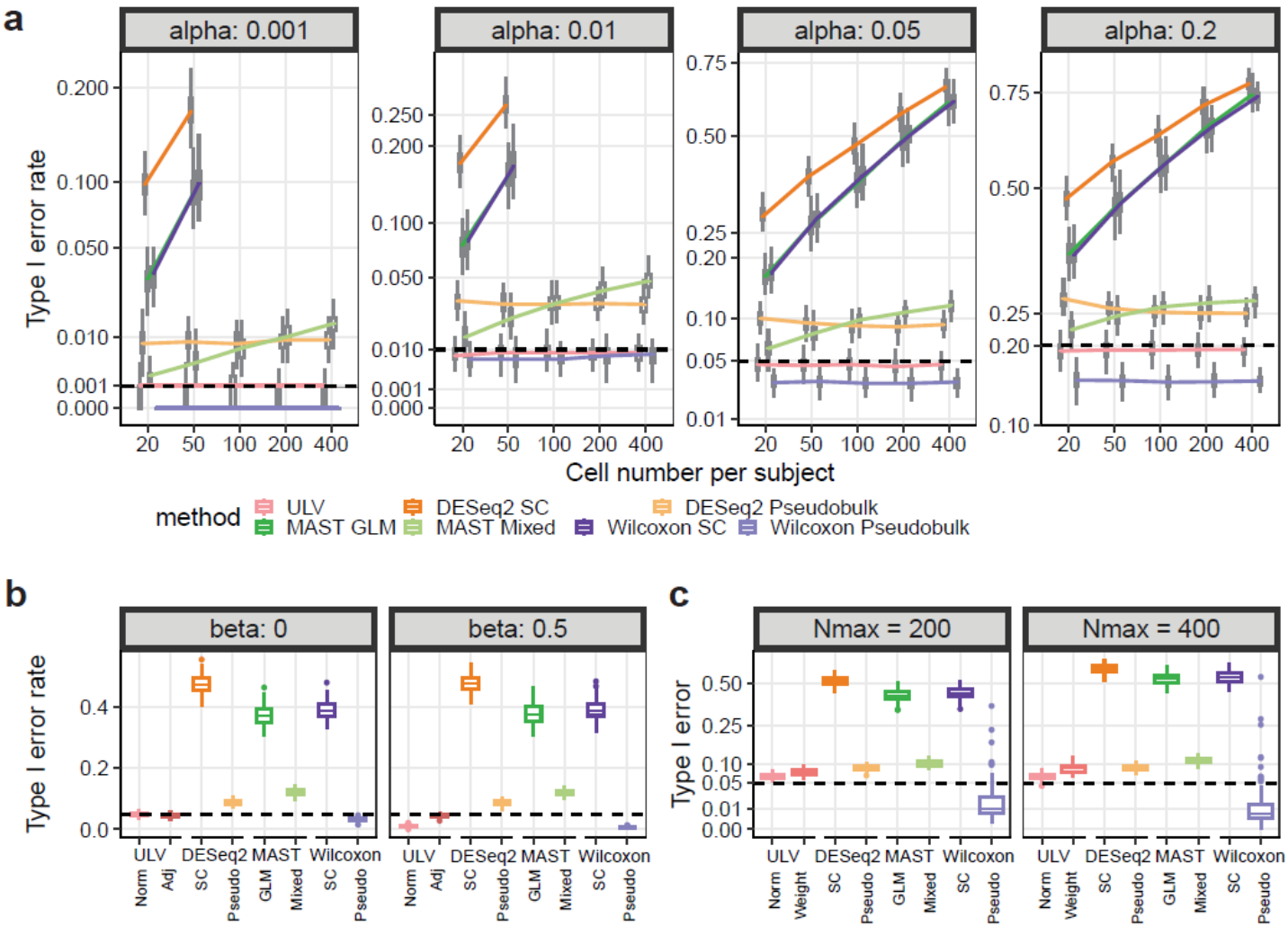
Appendix 3: Simulations

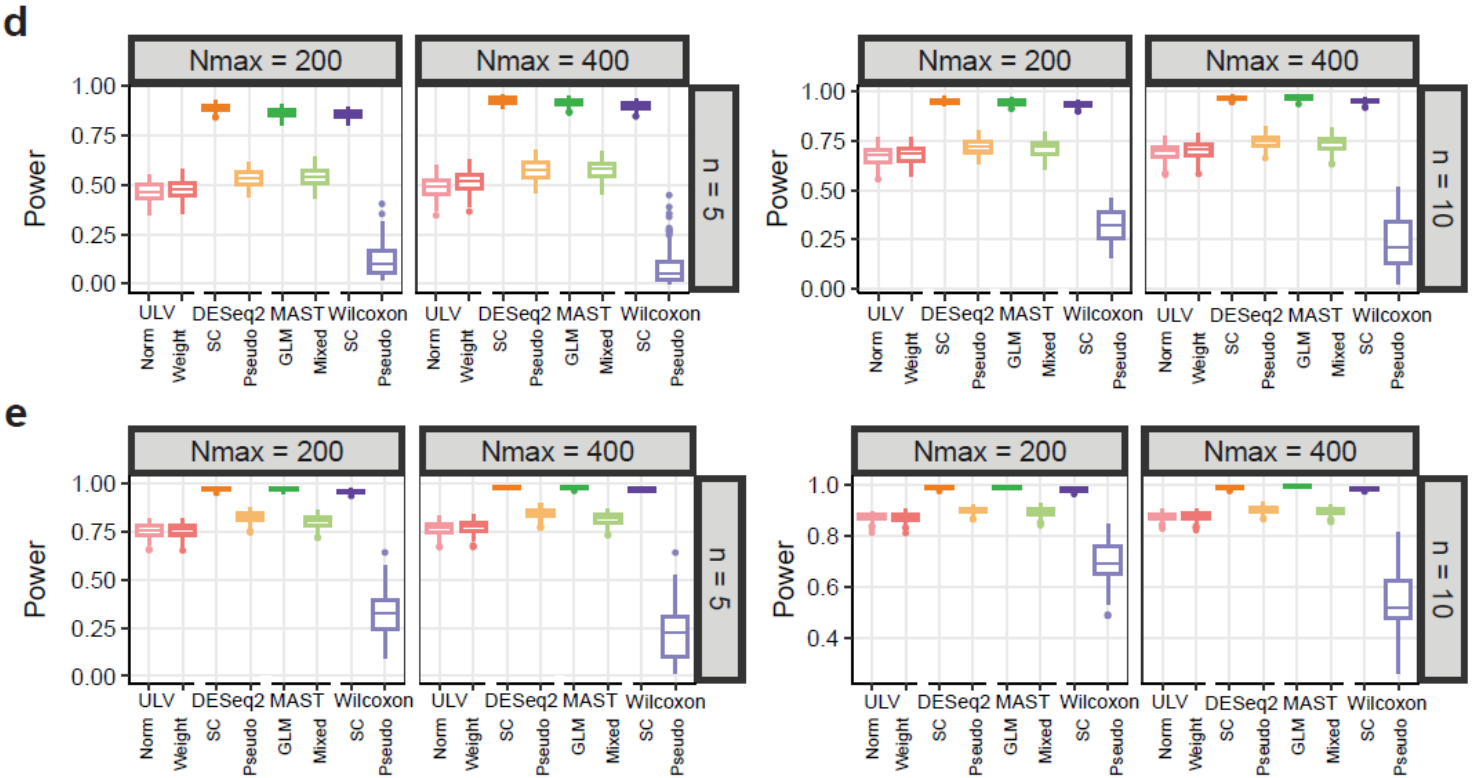
- Reference gene expression levels, inter/intraclass correlations, and correlations between genes were obtained using real data
- Simulation parameters
  - Number of cells per subject: 10-400
  - Number of subjects per group: 3-10 per group
  - Fold change: 1 to evaluate type I error rates, 4 to evaluate power
- Whether to adjust for covariates or unbalanced cluster sizes

Methods to Evaluate

- ULV
- Two DeSeq2 methods: assumes negative binomial distribution.
  - Method 1: treat cells as independent observations;
  - Method 2: pseudobulk: treat each subject as a bulk
- Two MAST methods: based on two-part, generalized linear models adapted for zero-inflated data.
  - Method 1: GLM;
  - Method 2: GLMM
- Two Wilcoxon methods

Simulation Results




d
e
Nmax = 200
Nmax = 400
n = 5
n = 10
Power
ULV
DESeq2
MAST
Wilcoxon
Norm
Weight
SC
Pseudo
GLM
Mixed

Appendix 4: Discussion of ULV

- Website: https://github.com/yu-zhaoxia/ULV
- Evaluate whether other difference metrics, such as the Hodges-Lehmann statistic, are better choices
- Extend the current work to model longitudinal data
- Develop robust methods for spatial single-cell data
- Extend the current work to perform causal inference