Title: Why do cell-level test methods for detecting differentially expressed genes fail in single-cell RNA-seq data?

Hanbin Lee[1,4]* and Buhm Han[1,2,3,4]*

1. Department of Medicine, Seoul National University College of Medicine, Republic of Korea

2. Department of Biomedical Sciences, BK21 Plus Biomedical Science Project, Seoul National University College of Medicine, Seoul, Republic of Korea

3. Interdisciplinary Program in Bioengineering, Seoul National University, Seoul, Republic of Korea

4. Genealogy Inc.

*** Corresponding Authors**

Hanbin Lee

Department of Medicine

Seoul National University College of Medicine

103 Daehak-ro, Jongno-gu, Seoul 03080, South Korea

TEL: 82-2-3668-7618

FAX: 82-2-741-0253

Email: hanbin973@snu.ac.kr

Buhm Han

Department of Biomedical Sciences

Seoul National University College of Medicine

103 Daehak-ro, Jongno-gu, Seoul 03080, South Korea

TEL: 82-2-3668-7618

FAX: 82-2-741-0253

Email: buhm.han@snu.ac.kr

Abstract

Earlier studies have found that cell-level tests for detecting differentially expressed genes (DEGs) exhibit inflated type 1 error in multi-subject scRNA-seq data. Pseudoreplication – cells from a same subject are non-independent observations – is thought to be the major cause which motivates the use of the pseudobulk. In this work, we argue that distributional misspecification, rather than pseudoreplication, might be a major cause of the inflated type 1 error in scRNA-seq DEG. We propose a simple solution to existing cell-level test methods that makes them robust against distributional misspecifications. Using real data, we show that the robust generalized linear models (GLMs) that are free of distribution constraints perform well, providing both controlled type 1 error and high statistical power. Since the solution is widely available in existing software, it can be immediately adapted to scRNA-seq analysis ecosystems.

Introduction

A straightforward approach for differentially expressed gene (DEG) analysis in scRNA-seq is to apply cell-level test statistics such as *t*-test and negative-binomial (NB) regression. However, studies found that these approaches produce overtly optimistic p-values leading to inflated type 1 error rates, especially in multi-subject studies[1-3]. Pseudoreplication – cells from the same subject are non-independent observations – has been considered as the major cause of the problem. Previously suggested works focused on dependence between cells within a subject (hereafter, subject autocorrelation). Pseudobulk, a widely adopted approach, aggregates the cells' expression within a subject (through mean or summation) and compares the expression across subjects[1]. Mixed models model the intra-subject dependence by including random intercepts[4].

In this work, rather than focusing on subject autocorrelation, we bring attention to the distributional misspecification of test statistics. Many tests rely on certain constraints on the underlying distribution, and we outline a few here. Likelihood ratio tests (LRTs) of generalized linear models (GLMs) such as Poisson regression or negative-binomial (NB) regression assume the corresponding probability distributions, respectively[5]. However, in single cell RNA data, these constraints are likely to be violated due to many factors, for example subject-specific batch effects (also a possible source of subject autocorrelation), zero inflation, and the presence of subpopulations of a particular cell type, to name a few[6,7]. Consequently, the standard error of the estimator is downward biased leading to inflated false discoveries (although upward bias is also theoretically possible).

Fortunately, there is a simple fix for such statistical tests. Furthermore, this fix is already widely implemented in standard software. Instead of using the conventional standard error derived under the principle of maximum likelihood estimator (MLE), using a robust standard error (also known as the sandwich estimator) derived from the theory of distributional misspecification can overcome this problem[8-10]. Most conventional GLMs implemented in the current software packages can be salvaged through this approach.

Using real data, we show that the robust versions of conventional test methods calibrate the false discoveries accurately without loss of power. Our work highlights that although subject autocorrelation cannot be ruled out completely as the cause of type 1 inflation, using the robust test statistics against

distributional misspecification can be a simple, practical, and widely applicable solution to obtain a reasonably calibrated result in multi-subject single-cell DEG analysis.

Results

In single-cell DEG, cells of a particular cell-type are isolated (**Figure 1a**) and their gene expression is subsequently compared (**Figure 1b**). Previous study, however, treating each cells as independent observations has led to inflated false discoveries (**Figure 1c**). This failure of cell-level test statistics in scRNA-seq DEG has been largely attributed to pseudoreplication (**Figure 1d**). We show that, distributional misspecification, rather than subject autocorrelation due to pseudoreplication, might be a major source of inflated type 1 error of cell-level methods (**Figure 1e**). We first outline the concept of robust standard errors[11,12]. In particular, we focus on generalized linear models (GLMs). GLM is generally considered as a maximum likelihood estimator (MLE) where the outcome follows a distribution from an exponential family conditional on the explanatory variables. However, in real data, it is hard to know whether the data truly follows the assumed distribution. In scRNA-seq data, even if we compare cells from the same group, such groups are still likely to be a heterogeneous collection of subtypes. As a result, for example, the expression might be multi-modally distributed which follows neither Poisson nor NB (**Figure 1e**).

Fortunately, results from classic statistics literature guarantee that the point estimates (or log-fold change in DEG context) are correct despite the misspecification and only the standard error should be modified[8-10]. Specifically, when the distribution is misspecified, the variance of the estimator is no more the inverse of the Fisher's information matrix but what is called the robust variance (or its square root, the standard errors)[12]. Implementation of robust standard error (robust SE) is widely available in all R and Python base environments although additional packages can deliver computationally more efficient implementations (see **Supplementary Table 1**)[12-14]. Henceforth, we denote test statistics of GLMs with the robust SE as the robust GLM test. The corresponding vanilla test will be called MLE GLM. Nevertheless, robust GLM still requires the observations to be independent, hence, is not robust to subject autocorrelation.

To demonstrate the well-calibrated nature of robust GLM tests, we conducted a null simulation based on real data with small (n=8, Kang et al.) and large (n=259, Nathan et al. and n=981, Yazar et al.) numbers of subjects[15-17] (**Figure 1f**). For each data, we re-assigned the case-control label (or the treatment status) randomly to cells. This way, we were able to perform a simple but realistic null simulation while preserving the structure of the original data, especially the subject autocorrelation (see

Methods).

In this benchmark, we compared two GLM methods (Poisson and NB regression, both MLE and robust), pseudobulk methods (edgeR[18] and DESeq[19]), and mixed models (Poisson and NB generalized linear mixed models or GLMMs) (**Figure 2**). As expected, all MLE GLMs performed badly, exhibiting deflated P-values. The robust counterparts produced well-calibrated P-values. Pseudobulk methods showed mixed results depending on the number of subjects. While DESeq2 was consistently conservative, edgeR was conservative with small number of subjects but suffered from inflated false discoveries with large number of subjects. Poisson GLMM also produced deflated p-values while NB GLMM performed well. As a summary of these results, we performed Kolmogorov-Sminorv (KS) test on the p-values respect to the uniform distribution on [0,1] (**Figure 2 and 3a**).

This result questions the contribution of subject autocorrelation to deflated p-values in scRNA-seq DEG. Essentially, robust GLMs require independence across observations and do not account for subject autocorrelation. Nevertheless, they performed well. In contrast, two pseudobulk methods (edgeR and DESeq2) and Poisson GLMM performed badly although they explicitly tackle subject autocorrelation (**Figure 3a**). To further investigate this issue, we applied robust GLMs to pseudobulk (hereafter, robust pseudobulk) data. In the smallest dataset (Kang et al., N=8), all pseudobulk methods did not perform well. This was expected because the number of observations in pseudobulk methods is simply the number of samples. When the sample size was large (Nathan et al. and Yazar et al.), like the single-cell case, robust GLMs returned better calibrated P-values compared to the MLE counterparts. This suggested that something other than autocorrelation might be the major cause of uncalibrated p-values, as we can consider the pseudobulk strategy to be robust against autocorrelation. Thus, tests robust to distributional assumptions, regardless of whether it accounts subject autocorrelation of not, returned reliable results in terms of type 1 error. Another modification of GLMs that can account subject autocorrelation is the cluster-robust standard errors given that the number of subjects is sufficiently large to guarantee asymptotic convergence[11,20]. In both Nathan et al. and Yazar et al. datasets, robust GLMs and cluster-robust GLMs were concordant (**Figure 3c**). The fact that the robust methods without considering autocorrelations and the robust methods considering autocorrelations gave the same results suggested that the impact of subject autocorrelation was ignorable for robust GLMs.

Next, we compared the power of the proposed methods (**Figure 4**, see **Methods**). We used the treatment label of the original studies of Kang et al. (interferon stimulation) and Nathan et al (tuberculosis infection). Yazar et al. was excluded because only healthy participants were included in the study. In this benchmark, cell-level methods that failed to calibrate type 1 error (KS test P-value<0.01/8=0.00125) were excluded leaving robust GLMs, robust pseudobulk and NB mixed model. Additionally, pseudobulk with edgeR and DESeq2 were compared, as they are widely used. When compared to robust Poisson regression, robust NB regression gave nearly identical results. Pseudobulk methods were relatively poorer than robust Poisson when the number of subjects was small but became comparable as the subjects increased. Notably, pseudobulk with robust GLMs performed better than edgeR and DESeq2. NB GLMM was comparable to robust Poisson, but the relative performance declined with larger number of subjects.

Discussion

In this work, we suggest that the distributional assumptions underlying the cell-level test statistics for DEG might be a major source of inflated false discoveries in single-cell DEG. Although subject autocorrelation has been considered as the major contributor, our result tells a different story. Although the role of subject autocorrelation cannot be fully ruled out with the current evidence, our study shows that cell-level test-statistics are well-calibrated for scRNA-seq DEG as long as the robust SE is used.

Robust GLMs, when directly applied to cells or to pseudobulks, provide calibrated type 1 error rates at the nominal level while achieving maximal power. Especially, when directly applied to cells, robust Poisson and NB performed best regardless of the number of subjects while pseudobulk had problems with small number of subjects. Unlike mixed models, additional computational cost of using robust GLM instead of MLE is minimal as the point estimate is retained. In our example, GLMMs were up to three orders of magnitude slower compared to robust GLMs. Hence, the robustness comes at nearly no cost making it an attractive choice.

Software availability is a crucial element for adoption of a method. Robust MLE tests can be performed easily with any programming languages (R and Python). Faster implementations are also available with additional libraries installed[12,21]. In **Supplementary Table 1** and an accompanied paper repository (https://github.com/hanbin973/DEGpaper), we provided a list of several such implementations and guides for the users.

In summary, widely adopted cell-level methods can be salvaged with minimal effort in terms of both implementation and computational cost through robust GLMs and their associated robust SEs. As long as the distributional violations are properly accounted, subject autocorrelation does not seem to breakdown the cell-level tests. Also, when the user chose to use pseudobulk, robust GLMs are recommended. Therefore, robust GLM can likely be the method of choice in scRNA-seq DEG.

Methods

MLE and robust GLMs

Throughout the benchmark, the same implementation was used. The commands used can be found in the paper repository (https://github.com/hanbin973/DEGpaper). Python statsmodels was used for Poisson regression, both MLE and robust[13]. MLE NB regression was performed with statsmodels, but the glmGamPoi was used instead for the robust version[14]. The reason is that unlike Poisson regression, negative-binomial regression requires more careful implementation as we discuss here briefly.

Robust Poisson regression only requires the conditional mean to be correctly specified.

$$G\left(E\left[y_{cg}\middle|X_c\right]\right) = X_c\beta_g$$

where X_c is the c-th row of the design matrix $\boldsymbol{X}$ and G is the link function. y_{cg} is the expression level of gene g of cell c. Note that the assumption is always true when X_c is an indicator for treatment groups. G is the log function for Poisson regression and NB regression. As a result, the fitting algorithm for the point estimate $\widehat{\beta_g}$ is identical to the MLE case. However, NB regression requires an additional condition related to the variance of y_{cg} conditional on X_{cg}[10,22]. As a result, MLE estimation algorithm becomes inadequate. glmGamPoi implements a correct algorithm for estimating β_g[14,23]. An initial estimate $\widehat{\beta_{g,0}}$ is first obtained following overdispersion estimation using $\widehat{\beta_{g,0}}$. The final estimate $\widehat{\beta_g}$ is then obtained using the overdispersion parameter obtained in the previous step.

The estimate for the MLE variance of $\widehat{\beta_g}$, $\widehat{\boldsymbol{Var}}(\widehat{\beta_g})$ is the inverse Hessian which is

$$\widehat{\boldsymbol{Var}}(\widehat{\beta_g}) = \left(\boldsymbol{X}'\boldsymbol{W}_g\boldsymbol{X}\right)^{-1}$$

where $\boldsymbol{W}_g$ is a diagonal matrix where the c-th diagonal value is $E\left[y_{cg}\middle|X_c\right]$ for log-link function. The apostrophe sign $'$ means the transpose of a matrix. The robust variance estimator is called the sandwich estimator:

$$\widehat{\boldsymbol{Var}}(\widehat{\beta_g}) = \left(\boldsymbol{X}'\boldsymbol{W}_g\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{\Sigma}_g\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{W}_g\boldsymbol{X}\right)^{-\mathbf{1}}$$

where $\boldsymbol{\Sigma}_g$ is a diagonal matrix where the c-th diagonal value is the square of $\epsilon_{cg} = Y_{cg} - G^{-1}\left(X_c\beta_g\right)$. To explain the name, $\left(\boldsymbol{X}'\boldsymbol{W}_g\boldsymbol{X}\right)^{-1}$ is called the bread and $\boldsymbol{X}'\boldsymbol{\Sigma}_g\boldsymbol{X}$ is called the meat. Note that these

estimators (both MLE and robust) are asymptotically valid provided $n_c \to \infty$ (the number of cells) and the observations are independent.

The cluster-robust variances estimator takes the form

$$\widehat{\boldsymbol{Var}}(\widehat{\beta_g}) = \left(\boldsymbol{X}'\boldsymbol{W}_g\boldsymbol{X}\right)^{-1}\left(\sum_{s=1\ to\ n_s} \boldsymbol{X}_s'\boldsymbol{\Sigma}_{s,g}\boldsymbol{X}_s\right)\left(\boldsymbol{X}'\boldsymbol{W}_g\boldsymbol{X}\right)^{-\mathbf{1}}$$

where $\boldsymbol{X}_s$ is the subset of the design matrix $\boldsymbol{X}$ that belongs to subject s and

$$\left[\boldsymbol{\Sigma}_{s,g}\right]_{i,j} = \epsilon_{ig}\epsilon_{jg}$$

for cell i,j belonging to subject s[12]. This estimator is valid asymptotically provided $n_s \to \infty$. As described in the main text, cluster-robust SE is robust to subject autocorrelation. Only cluster-robust Poisson (no NB) was considered in this work because glmGamPoi did not support cluster-robust inference.

Pseudobulk methods

We benchmarked two pseudobulk methods, edgeR and DESeq2[18,19]. The pseudobulk count matrix was generated by summing the UMI counts within each replicate. The replicates were defined for each subject and treatment label.

Mixed models

We used Poisson and NB GLMM implemented by the glmmTMB library in R[4]. The subject-level random intercept was included for each gene-wise regression.

Type 1 error benchmark

For each dataset, we randomly selected 5,000 cells. The reason for the downsampling was due to the large computational burden of mixed models. We found that while GLMs take approximately 0.002 seconds per gene, GLMMs implemented in glmmTMB took more than 3 seconds under this environment. Another standard library, lme4 (implemented in R), took even longer (similar with Poisson GLMM and ~10 times slower with NB GLMM).

After subsampling the cells, the treatment label was assigned at random with probability 0.5. Next, the

significance and the log-fold change of the treatment label was tested across all genes that had at least 10 counts. The Kolmogorov-Sminorv (KS) test was performed using the 'scipy.stats.kstest' function. The code can be found in the paper repository (https://github.com/hanbin973/DEGpaper).

Datasets

Kang et al. is a peripheral mononuclear blood cell (PBMC) dataset derived from 8 subjects[15]. PBMCs were divided into two groups where only one group was stimulated using interferon. Total 13,836 cells were obtained from 10X Chromium platform (Accession code: GSE96583). Nathan et al. dataset was obtained from 259 subjects with previous history of tuberculosis infection[16]. 500,089 T-cells were isolated using the CITE-seq technology (Accession code: GSE158769). Yazar et al. dataset was obtained from 982 healthy donors consisting of total 1,248,980 PBMCs[16,17]. The data was generated using the 10X Chromium platform (obtained from https://cellxgene.cziscience.com/).

Code availability

The analysis code used in this paper can be found at https://github.com/hanbin973/DEGpaper.

Conflict of interest

Buhm Han is the CTO of the Genealogy Inc.

Figure legends

Figure 1. Overview of the study. a. cells are isolated by cell-type **b.** gene expression of cells from cases are compared to those from the controls. **c.** a hypothetical volcano plot where majority of the genes are found to be differentially expressed. **d.** subjects of a study are independent observations while cells within each subject are non-independent each other. **e.** the true distribution (dotted line) of gene expression can be different from the assumed model (solid line)**. f.** the null simulations compare the P-value distribution of a statistical test to the expected distribution under the null hypothesis.

Figure 2. Quantile-Quantile (QQ) plot of P-values of the methods respect to the uniform distribution on [0,1] in log-scale. a. MLE GLMs, robust GLMs, pseudobulks (edgeR and DESeq2) and mixed models (Poisson and NB GLMM) applied to the Kang et al. dataset. **b.** identical to **a.** but using the Nathan et al. dataset. **c.** identical to **a.** but using the Yazar et al. dataset.

Figure 3. Further evaluation of scRNA-seq DEG methods. a. KS test results of MLE GLMs, robust GLMs, pseudobulks (edgeR and DESeq2) and mixed models (Poisson and NB GLMM). **b.** MLE GLMs and robust GLMs applied to pseudobulk data of Kang et al., Nathan et al. and Yazar et al. datasets **c.** QQ plot of cluster-robust Poisson regression applied to Nathan et al. and Yazar et al. datasets.

Figure 4. Power comparison of methods. a. QQ plot of P-values of pseudobulk (edgeR, DESeq2, robust GLMs), robust NB and NB GLMM applied to the Nathan et al. dataset **b.** same as a. but applied to the Yazar et al. dataset.

References

1 Zimmerman, K. D., Evans, C. & Langefeld, C. D. Reply to: A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. *Nat Commun* **13**, 7852, doi:10.1038/s41467-022-35520-x (2022).

2 Zimmerman, K. D., Espeland, M. A. & Langefeld, C. D. A practical solution to pseudoreplication bias in single-cell studies. *Nat Commun* **12**, 738, doi:10.1038/s41467-021-21038-1 (2021).

3 Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* **12**, 5692, doi:10.1038/s41467-021-25960-2 (2021).

4 Brooks, M. E. *et al.* glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal* **9**, doi:10.32614/rj-2017-066 (2017).

5 Dobson, A. J. & Barnett, A. G. *An introduction to generalized linear models, fourth edition*. 4 edn, (CRC Press).

6 Sarkar, A. & Stephens, M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat Genet* **53**, 770-777, doi:10.1038/s41588-021-00873-4 (2021).

7 Gorin, G., Vastola, J. J., Fang, M. & Pachter, L. Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments. *Nat Commun* **13**, 7620, doi:10.1038/s41467-022-34857-7 (2022).

8 Huber, P. J. in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* Vol. 1: Statistics 221-233 (University of California Press, 1967).

9 White, H. Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **50**, doi:10.2307/1912526 (1982).

10 McCullagh, P. Quasi-Likelihood Functions. *The Annals of Statistics* **11**, doi:10.1214/aos/1176346056 (1983).

11 Arellano, M. PRACTITIONERS' CORNER: Computing Robust Standard Errors for Within-groups Estimators*. *Oxford Bulletin of Economics and Statistics* **49**, 431-434, doi:10.1111/j.1468-0084.1987.mp49004006.x (2009).

12 Zeileis, A., Köll, S. & Graham, N. Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R. *Journal of Statistical Software* **95**, doi:10.18637/jss.v095.i01 (2020).

13 Seabold, S. & Perktold, J. in *9th Python in Science Conference.* (2010).

14 Ahlmann-Eltze, C. & Huber, W. glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data. *Bioinformatics* **36**, 5701-5702, doi:10.1093/bioinformatics/btaa1009 (2021).

15 Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol* **36**, 89-94, doi:10.1038/nbt.4042 (2018).

16 Nathan, A. *et al.* Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature* **606**, 120-128, doi:10.1038/s41586-022-04713-1 (2022).

17 Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041, doi:10.1126/science.abf3041 (2022).

18 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

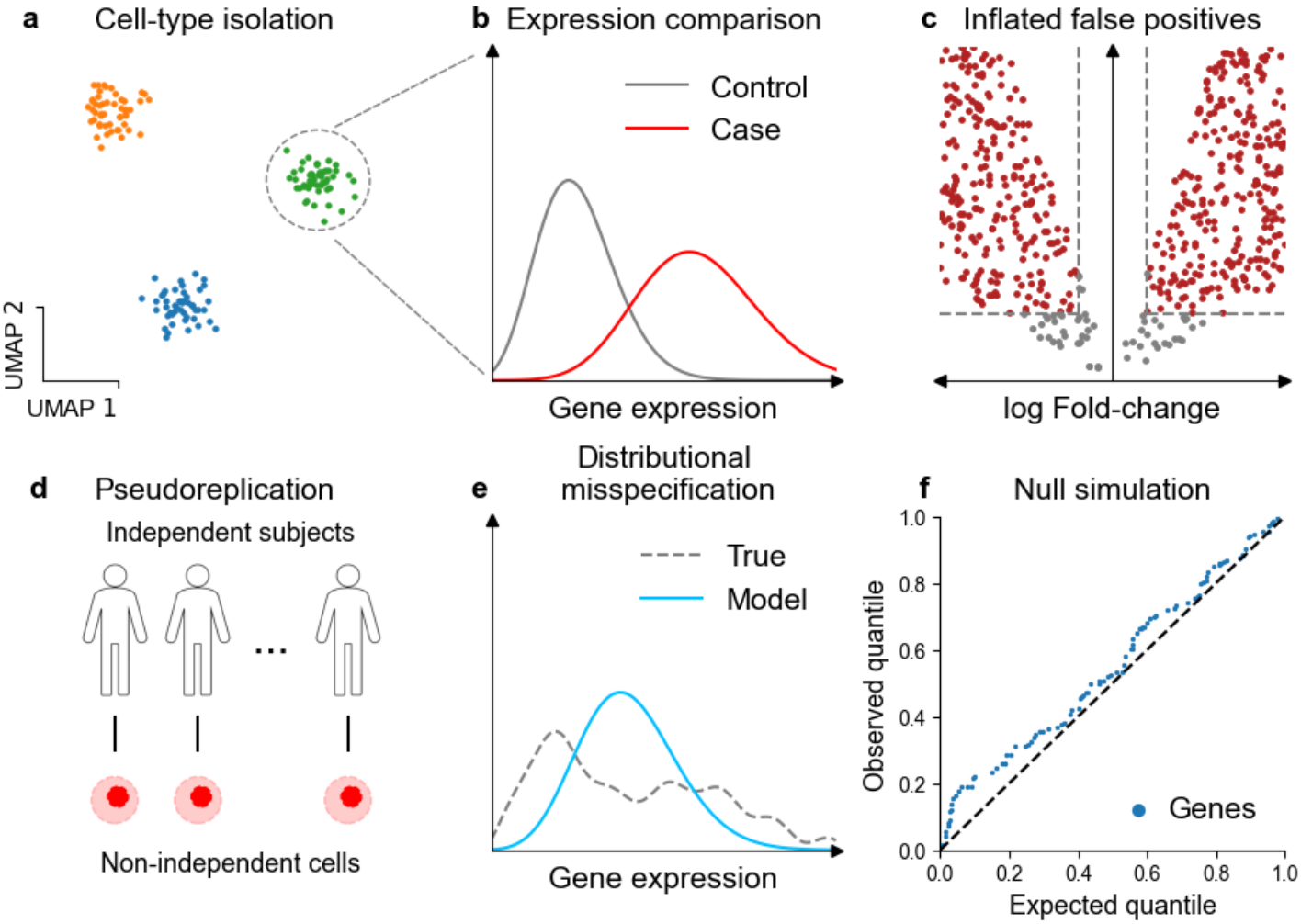
19 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

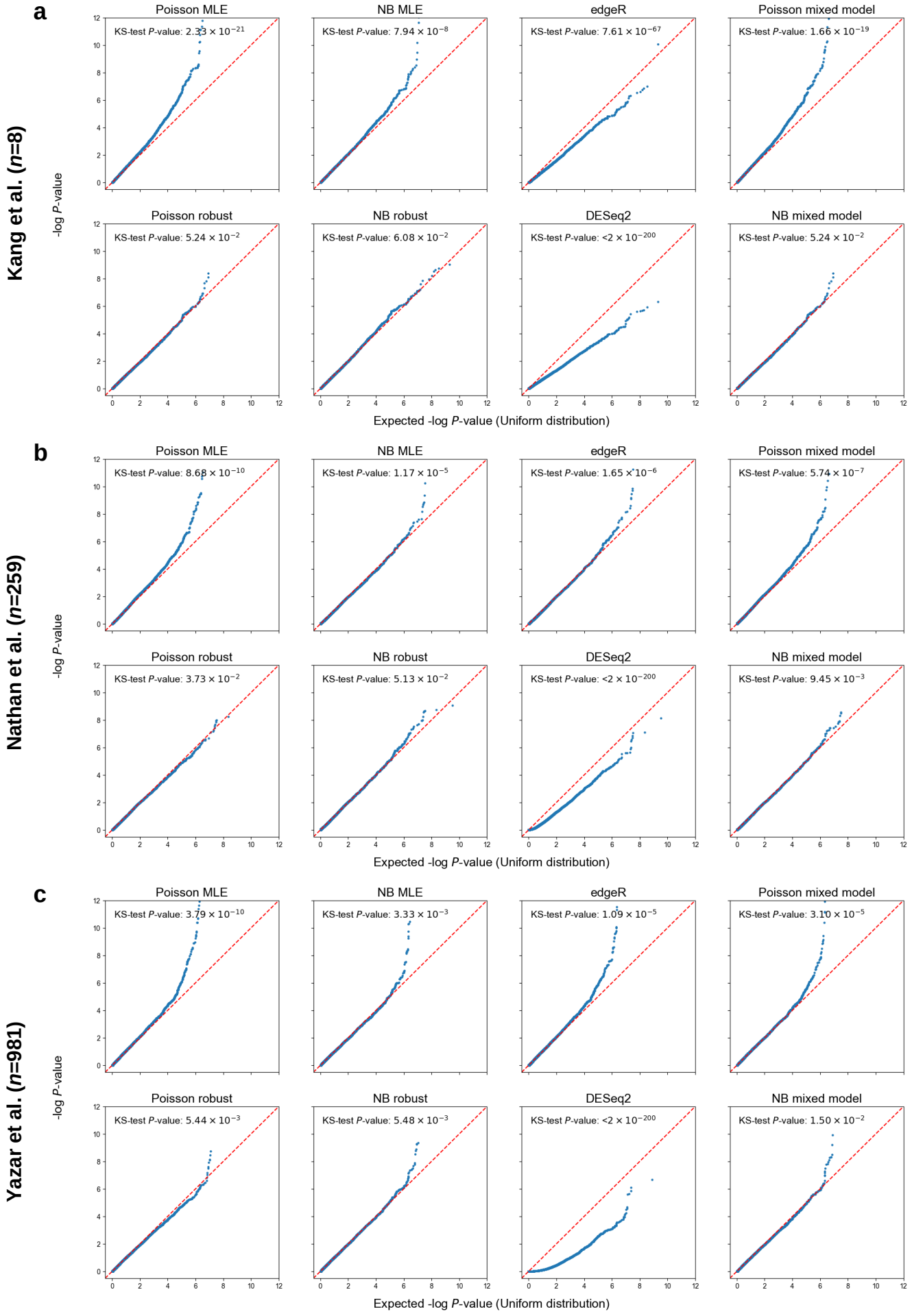
20 Liang, K.-Y. & Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22, doi:10.1093/biomet/73.1.13 (1986).

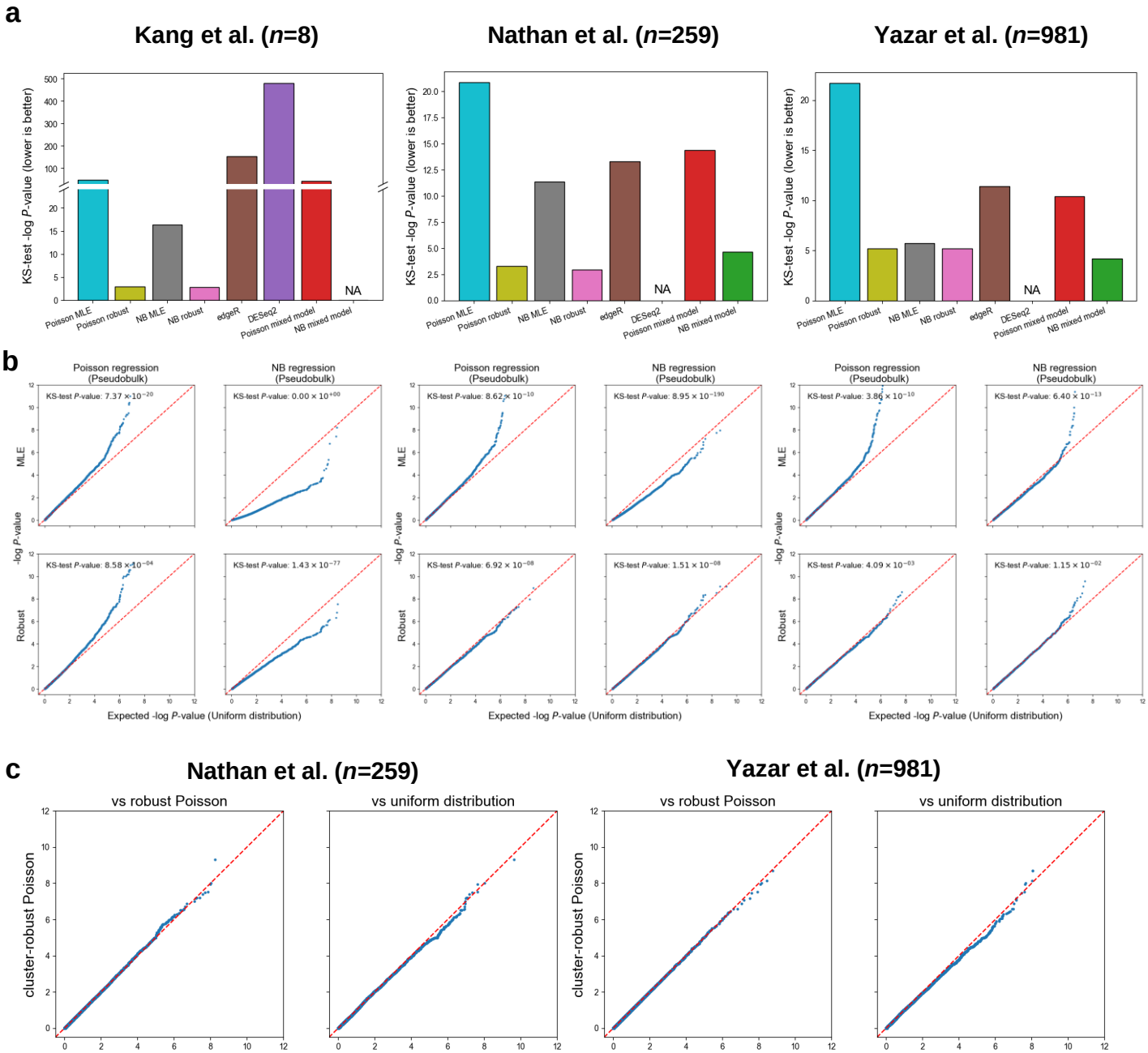
21 Bergé, L. Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. (Department of Economics at the University of Luxembourg, 2018).

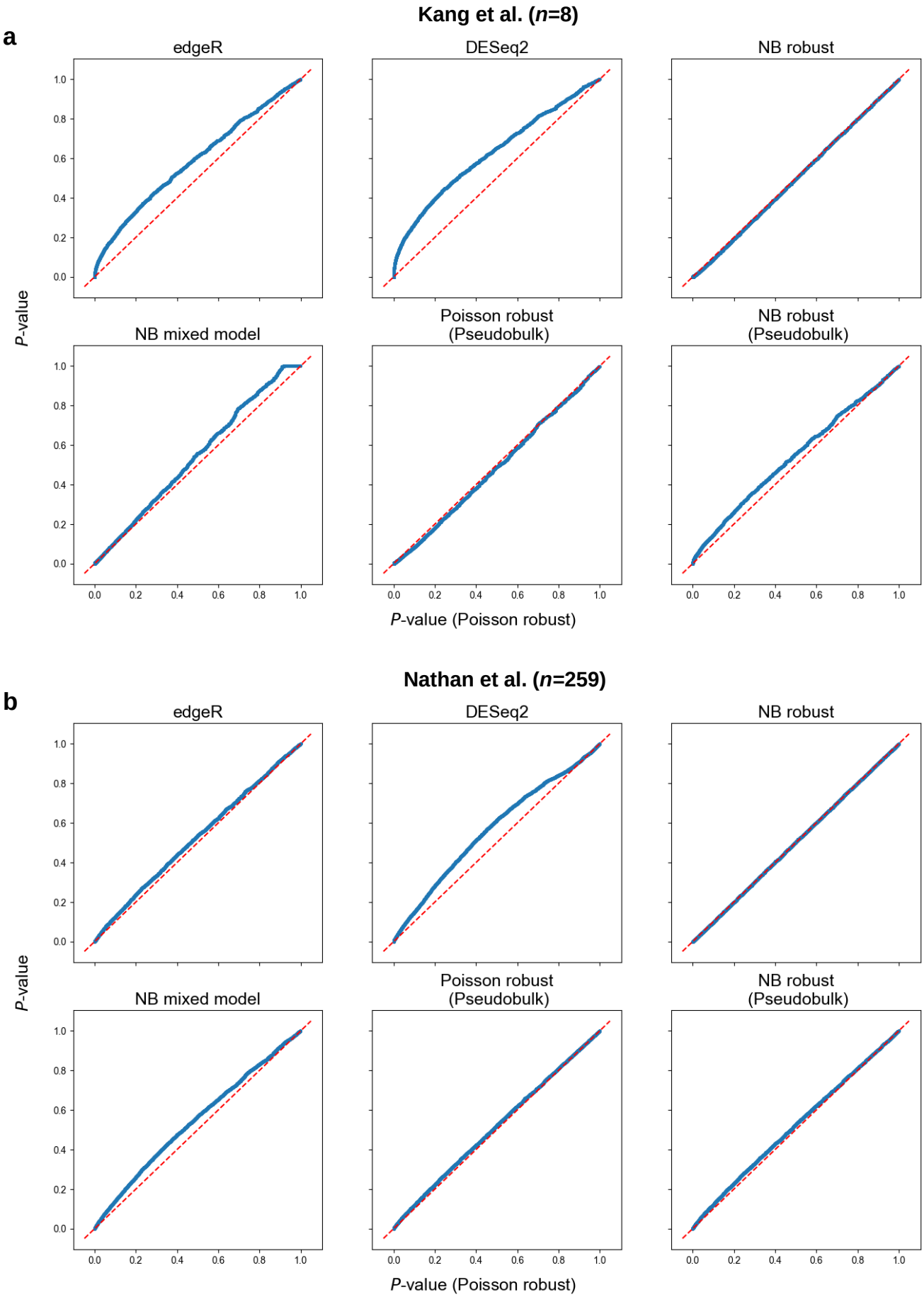
22 Wooldridge, J. M. in *Handbook of Applied Econometrics Volume 2: Microeconomics* 321-368 (1999).

23 Lund, S. P., Nettleton, D., McCarthy, D. J. & Smyth, G. K. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat Appl Genet Mol Biol* **11**, doi:10.1515/1544-6115.1826 (2012).


a
Cell-type isolation
UMAP 2
UMAP 1
b
Expression comparison
Control
Case
Gene expression
c
Inflated false positives
log Fold-change
d
Pseudoreplication
Independent subjects
...
Non-independent cells
e
Distributional misspecification
True
Model
Gene expression
f
Null simulation
Observed quantile
Expected quantile
Genes
0.0
0.2
0.4
0.6
0.8
1.0


a
Kang et al. (n=8)
Poisson MLE
KS-test P-value: 2.33 × 10^−21
NB MLE
KS-test P-value: 7.94 × 10^−8
edgeR
KS-test P-value: 7.61 × 10^−67
Poisson mixed model
KS-test P-value: 1.66 × 10^−19
Poisson robust
KS-test P-value: 5.24 × 10^−2
NB robust
KS-test P-value: 6.08 × 10^−2
DESeq2
KS-test P-value: <2 × 10^−200
NB mixed model
KS-test P-value: 5.24 × 10^−2
b
Nathan et al. (n=259)
Poisson MLE
KS-test P-value: 8.68 × 10^−10
NB MLE
KS-test P-value: 1.17 × 10^−5
edgeR
KS-test P-value: 1.65 × 10^−6
Poisson mixed model
KS-test P-value: 5.74 × 10^−7
Poisson robust
KS-test P-value: 3.73 × 10^−2
NB robust
KS-test P-value: 5.13 × 10^−2
DESeq2
KS-test P-value: <2 × 10^−200
NB mixed model
KS-test P-value: 9.45 × 10^−3
c
Yazar et al. (n=981)
Poisson MLE
KS-test P-value: 3.79 × 10^−10
NB MLE
KS-test P-value: 3.33 × 10^−3
edgeR
KS-test P-value: 1.09 × 10^−5
Poisson mixed model
KS-test P-value: 3.10 × 10^−5
Poisson robust
KS-test P-value: 5.44 × 10^−3
NB robust
KS-test P-value: 5.48 × 10^−3
DESeq2
KS-test P-value: <2 × 10^−200
NB mixed model
KS-test P-value: 1.50 × 10^−2
-log P-value
Expected -log P-value (Uniform distribution)


a
Kang et al. (n=8)
Nathan et al. (n=259)
Yazar et al. (n=981)
KS-test -log P-value (lower is better)
Poisson MLE
Poisson robust
NB MLE
NB robust
edgeR
DESeq2
Poisson mixed model
NB mixed model
NA
b
Poisson regression (Pseudobulk)
NB regression (Pseudobulk)
MLE
Robust
-log P-value
KS-test P-value: 7.37 × 10^−20
KS-test P-value: 0.00 × 10^+00
KS-test P-value: 8.58 × 10^−04
KS-test P-value: 1.43 × 10^−77
KS-test P-value: 8.62 × 10^−10
KS-test P-value: 8.95 × 10^−190
KS-test P-value: 6.92 × 10^−08
KS-test P-value: 1.51 × 10^−08
KS-test P-value: 3.86 × 10^−10
KS-test P-value: 6.40 × 10^−13
KS-test P-value: 4.09 × 10^−03
KS-test P-value: 1.15 × 10^−02
Expected -log P-value (Uniform distribution)
c
Nathan et al. (n=259)
Yazar et al. (n=981)
vs robust Poisson
vs uniform distribution
cluster-robust Poisson


a
Kang et al. (n=8)
edgeR
DESeq2
NB robust
NB mixed model
Poisson robust (Pseudobulk)
NB robust (Pseudobulk)
P-value
P-value (Poisson robust)
b
Nathan et al. (n=259)
edgeR
DESeq2
NB robust
NB mixed model
Poisson robust (Pseudobulk)
NB robust (Pseudobulk)
P-value
P-value (Poisson robust)