Title: On the cause of inflated type 1 error in single cell DEG

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 can be another, if not primary, cause of the inflated type 1 error in scRNA-seq DEG. We show that the robust sandwich estimator for distributionally misspecified maximum likelihood estimator (MLE) produces calibrated type 1 error rate. Using real data, we show that the robust method performs surprisingly well, providing both controlled type 1 error and high statistical power, and is superior to many other methods that explicitly account only for pseudoreplication. Since our suggested solution is already 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 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. For this reason, previous works focused only on accounting for dependence between cells within a subject (hereafter, subject autocorrelation). For example, Pseudobulk, a widely adopted approach, aggregates the cell's expression within a subject to remove autocorrelation and compares the expression across subjects[1]. Mixed models explicitly model the intra-subject dependence by including random intercepts[4].

In this work, we argue that another major, if not primary, cause of inflation is the distributional misspecification of test statistics. Currently used tests rely on certain constraints on the underlying distribution, as most are based on generalized linear models (GLMs) assuming Poisson or negative-binomial (NB) distributions[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]. These violations will only be exacerbated in multi-subject studies, obviously. Consequently, the standard error of the estimator can be downward biased leading to inflated false discoveries (although upward bias is also theoretically possible).

Fortunately, there exists a practical solution for the distributional misclassification. 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]. One good aspect of this solution is that it is already widely implemented in the standard software packages. Thus, most analyses using conventional GLMs implemented in the current software packages can be salvaged by updating a few lines of code.

Strikingly, applying robust sandwich estimator to multi-subject real data showed that it achieves better calibrated false discovery rate and power than other methods, including methods that explicitly

account for autocorrelation. This raises a question which is the primary cause of inflation of false positives: subject autocorrelation or distributional misspecification. Although both will practically exist in real data, our results support that accounting for the latter can be more crucial for obtaining a well calibrated DEG results. Our work highlights that using the robust test statistics against distributional misspecification can be a simple, practical, and widely applicable solution to control false positives in multi-subject single-cell DEG analysis while retaining maximum power.

RESULTS

Overview

In single-cell DEG analysis, cells of a particular cell-type are isolated (**Figure 1a**) and their gene expression is subsequently compared, often across multiple subjects (**Figure 1b**). However, a naïve approach treating each cell as independent observation can lead to inflated false discoveries (**Figure 1c**). This failure of cell-level test statistics in multi-subject scRNA-seq DEG has been largely attributed to pseudoreplication (**Figure 1d**). Here, we argue that distributional misspecification can be another major, possibly primary, source of inflated type 1 error of cell-level methods (**Figure 1e**).

Single-cell DEG test methods are generally built upon generalized linear models (GLMs). GLM assumes that the outcome follows a distribution from an exponential family conditional on the explanatory variables and comes with a maximum likelihood estimator. However, in real data, it is hard to know whether the data truly follows the assumed distribution. For example, in scRNA-seq data, even if we constrain our focus on a single cell-type, it can still be a heterogeneous collection of subtypes. As a result, the expression might be multi-modally distributed which follows neither Poisson nor NB (**Figure 1e**).

Fortunately, there exists a solution for distributional misspecification. Results from classic statistics literature suggests that only the standard error required modification[8-10], as the point estimates (or log-fold change in DEG context) are unaffected from the misspecification. 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 (often called, sandwich variance)[11]. Implementation of robust standard error (robust SE), which estimates the sandwich variance, is widely available in all R and Python base environments. Obtaining robust SE induces very little computational burden compared to obtaining the conventional SE, and additional packages can deliver computationally more efficient implementations (see **Supplementary Table 1**)[11-13].

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. Note that robust GLM still requires the observations to be independent, and hence, is theoretically not robust to subject autocorrelation.

Null simulations using multi-subject real datasets

To evaluate the performance of robust GLM tests compared to naïve (MLE) GLM tests, we conducted a null simulation using three real datasets with one small (*n*=8, Kang et al.) and two large (*n*=259, Nathan et al. and *n*=981, Yazar et al.) numbers of subjects[14-16] (**Figure 1f**). We wanted to design a realistic null scenario with minimal artificial assumptions. To this end, for each data, we randomly assigned the case-control label (or the treatment status) 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**). Since subjects were not originally paired in these datasets, the number of subjects was doubled by this randomization for pseudobulk methods (one case and one control simulated from one subject).

In this benchmark, we compared robust GLM to MLE GLM both assuming Poisson and NB regression (**Figure 2**). In addition, we also evaluated pseudobulk methods (edgeR[17] and DESeq[18] using pseudobulk mean expression), and mixed models (Poisson and NB generalized linear mixed models or GLMMs). As previously reported and expected, all MLE GLMs performed poorly, exhibiting deflated P-values (inflated false positives). However, the robust counterparts produced well-calibrated null P-values. When we performed Kolmogorov-Sminorv (KS) test on the P-values respect to the uniform distribution on [0,1], robust tests gave much larger P-values (closer to null) than MLE tests (**Figure 3a**). Pseudobulk methods showed mixed results depending on the number of subjects. DESeq2 was consistently overly conservative regardless of datasets. edgeR was conservative with small number of subjects but suffered from inflated false discoveries with large number of subjects. Poisson GLMM also produced inflated false positives for all three datasets. Excluding the robust tests, NB GLMM was the only method that performed reasonably well. However, NB GLMM did not converge for a small dataset (Kang et al., *n*=8).

This result is striking because robust GLM that tackles distributional misspecification worked very well while other methods that tackle autocorrelation did not perform well. This raises a question which is the primary source of P-value deflation: subject autocorrelation or distributional misspecification, even though both would practically exist in real data. Essentially, robust GLMs require independence across observations and do not account for subject autocorrelation, but they performed well. In contrast, two pseudobulk methods (edgeR and DESeq2) performed poorly although they can be considered to completely remove subject autocorrelation.

To further investigate this issue, we applied robust GLMs to pseudobulk data (hereafter, robust pseudobulk). In the smallest dataset (Kang et al., N=8), all pseudobulk methods, including robust pseudobulk, did not perform well (**Figure 3b**). This was expected because the number of observations in pseudobulk methods is simply the number of subjects. This emphasized that, as suggested previously[3], pseudobulk methods should only be used when the number of subjects is sufficiently large.

When the number of subjects was large (Nathan et al. and Yazar et al.), like the cell-level test cases above, robust GLMs returned better calibrated P-values compared to the MLE counterparts (**Figure 3b and 3c**). Therefore, tests robust to distributional assumptions, regardless of whether they account for subject autocorrelation or not, returned reliable results in terms of type 1 error. This observation suggests that something other than autocorrelation (likely distributional misspecification) might be the major cause of uncalibrated P-values, as we can consider the pseudobulk strategy to be robust against autocorrelation.

Another modification of GLMs that can account for subject autocorrelation is the cluster-robust standard errors that guarantee asymptotic convergence given that the number of subjects is sufficiently large[19,20]. In both Nathan et al. and Yazar et al. datasets, robust GLMs and cluster-robust GLMs were concordant (**Figure 3d**). 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 negligible when we applied robust GLMs.

Power comparison using real datasets

Comparing power of methods under a realistic setting is much more complex than null simulations. Although a simple randomization can achieve a reasonable null simulation, to evaluate power, we need many assumptions such as true DEGs and their effect sizes. To prevent these prior assumptions from affecting results, we employed an indirect approach that compares P-values between methods given a real dataset with original treatment label. Specifically, 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. If we assume that

DEGs with significant P-values were truly driven by the treatment, then we can indirectly “presume” that a method with smaller P-values is more powerful, conditioned on that nulls were well calibrated.

In this benchmark, cell-level methods that failed to calibrate type 1 error (KS test P-value<0.01/8=0.00125) were excluded. This filtering left us robust GLMs, robust pseudobulk, and NB mixed model. However, we included pseudobulk with edgeR and DESeq2 for our comparison, despite of not meeting the criteria, because they are widely used in the community.

Figure 4 shows the relative P-value comparison when compared to our robust Poisson regression. In this plot, if the line shows a bulge above the diagonal, it means P-values are larger (power is lower), and if the line shows a deep under the diagonal, it means P-values are smaller (power is higher). When compared to robust Poisson, two pseudobulk methods (edgeR and DESeq2) were less powerful than robust Poisson although edgeR improved with larger number of subjects. Robust NB and robust Poisson had identical power. NB GLMM showed comparable performance but was slightly less powerful than robust Poisson. Finally, pseudobulk with robust GLMs were nearly as powerful as the cell-level counterparts with improving performance with the increasing number of subjects.

Speed comparison

We compared the speed of the methods. Like the previous analysis, we excluded methods that failed type 1 error test, leaving us robust GLMs, robust pseudobulk, and NB mixed model. The robust pseudobulk can be instantly performed, as the number of subjects is much smaller than cell-level tests. When we compared cell-level tests, robust GLMs was >100 times faster than NB mixed model. For testing DEG for 10,000 genes in 100,000 cells, robust GLM took only about an hour while NB mixed model took 6.4 days (**Figure 5**).

Discussion

In this work, we propose the use of robust (sandwich) estimator for DEG analysis of multi-subject scRNA-seq study. Robust GLMs, both when directly applied to cells and when applied to pseudobulks, provided calibrated type 1 error rates while achieving maximal power. Especially, when directly applied to cells, robust Poisson and NB performed the best regardless of the number of subjects, while pseudobulk strategy was only applicable to a large number of subjects. Unlike mixed models that incur a large computational burden, additional computational cost of using robust GLM instead of MLE is minimal. In our example, robust GLMs were >100 times faster compared to mixed models. Hence, the robustness comes at nearly no cost making it an attractive choice.

The fact that our robust GLM works well in various contexts raised a question of which is the primary source of p-value deflation in cell-level tests involving multi-subjects. Our results suggest that the distributional assumptions underlying the cell-level test statistics for DEG might be the major source of inflated false discoveries in single-cell DEG. Although subject autocorrelation has been considered as the major contributor to date, our result tells a possibly 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.

Software availability is a crucial element for adoption of a method. Robust GLM tests can be performed easily with existing functions in popular programming languages (R and Python). Faster implementations are also available with additional libraries installed[11,21]. 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, even when the user chose to use pseudobulk, robust GLMs are still recommended. We expect that robust GLM can be widely used to reduce false positives and maximize power in future scRNA-seq studies.

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[12]. MLE NB regression was performed with statsmodels, but the glmGamPoi was used instead for the robust version[13]. 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.

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

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[13,23]. 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 $E\left[y_{cg}|X_c\right]$ 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

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

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[11]. 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[17,18]. 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[14]. 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[15]. 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[15,16]. 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). MLE GLMs and robust GLMs applied to pseudobulk data of **b.** Kang et al., **c.** Nathan et al. and **d.** Yazar et al. datasets **e.** 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.

Figure 5. Time comparison of methods. Runtime of three methods for 10,000 DEGs on 100,000 cells.

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 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).

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

13 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).

14 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).

15 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).

16 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).

17 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).

18 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).

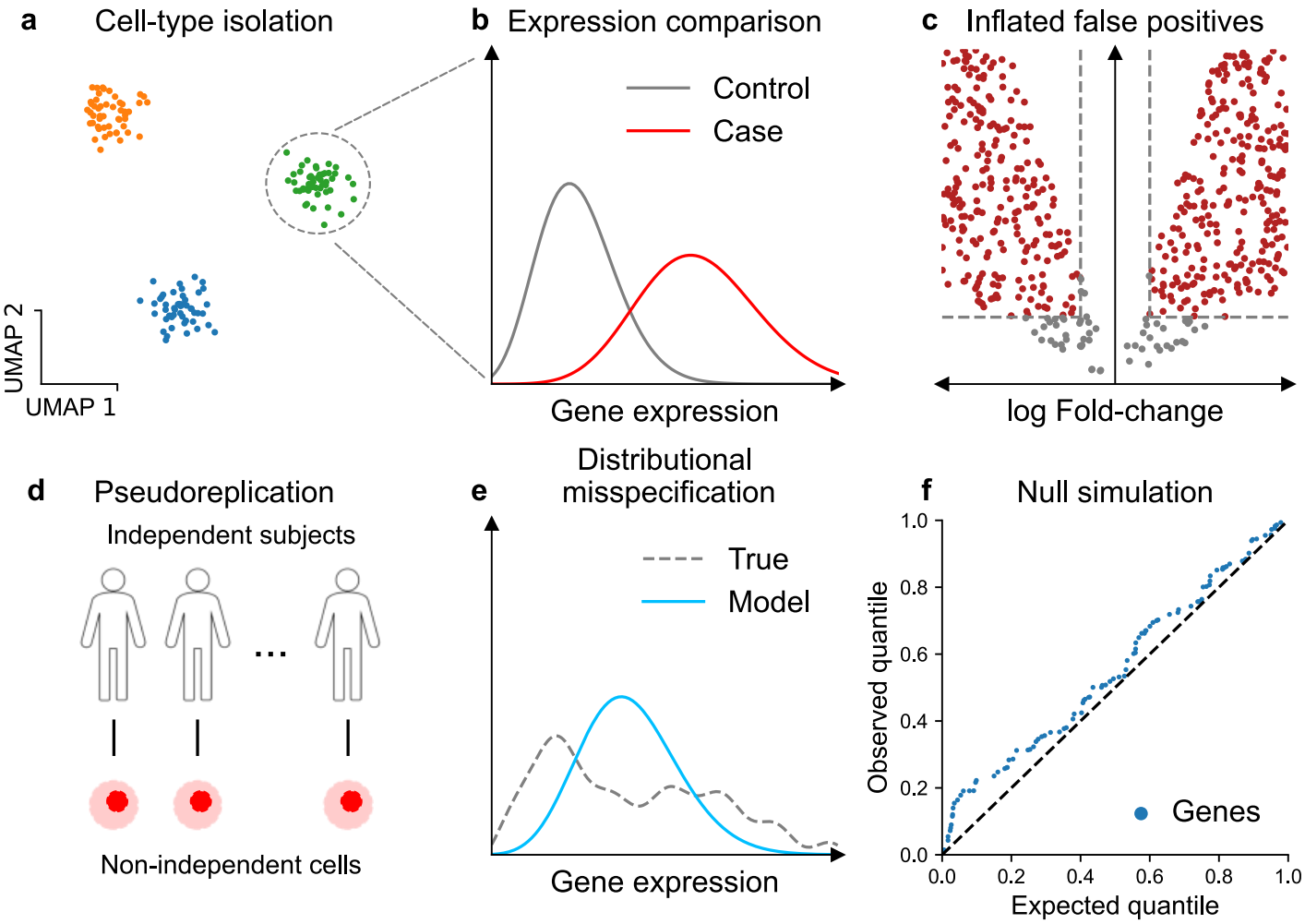
19 Liang, K.-Y. & Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22, doi:10.1093/biomet/73.1.13 (1986).

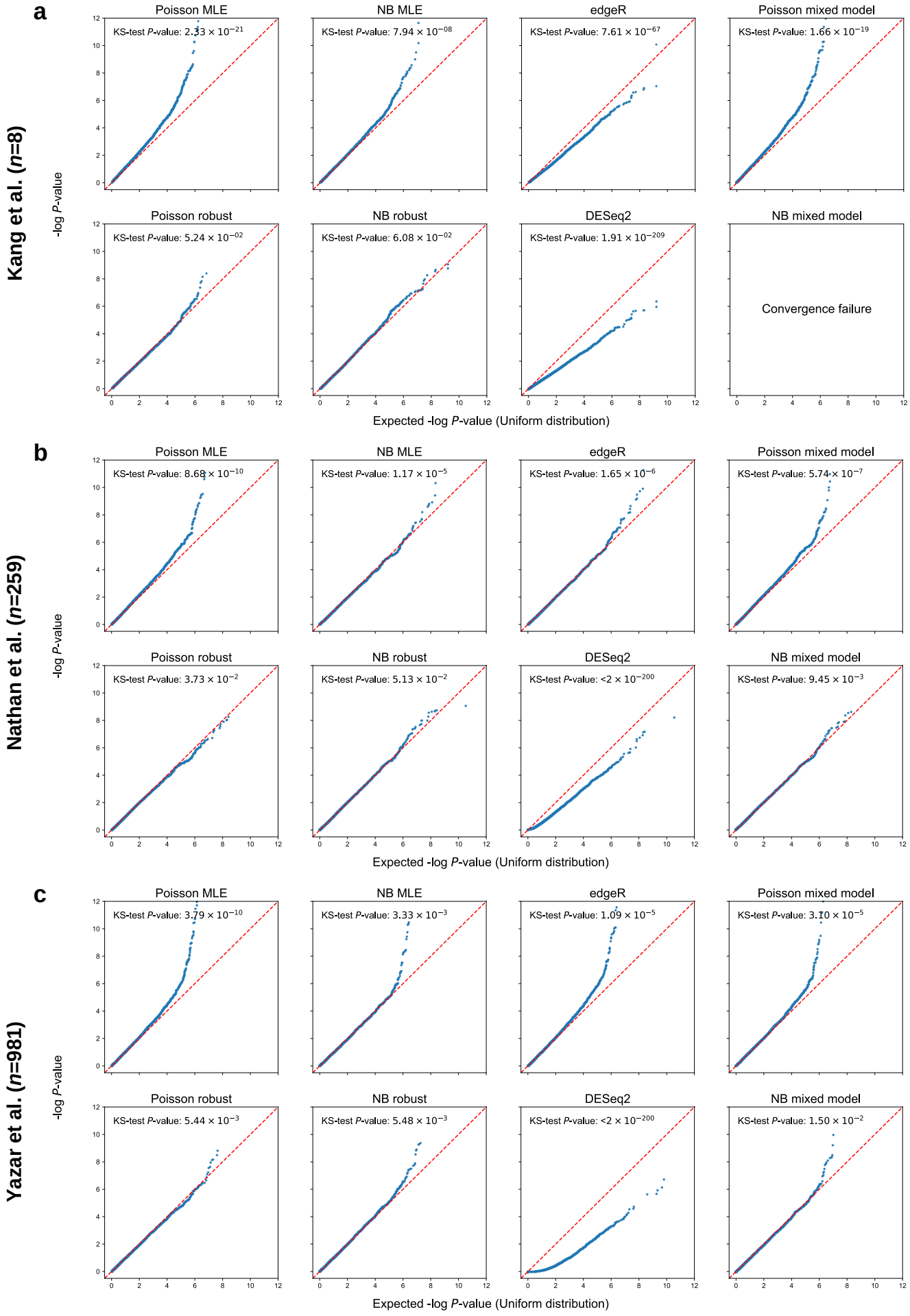
20 Arellano, M. PRACTITIONERS' CORNER: Computing Robust Standard Errors for Within-groups Estimators*. *Oxford Bulletin of Economics and Statistics* **49**, 431-434, doi:10.1111/j.1468-0084.1987.mp49004006.x (2009).

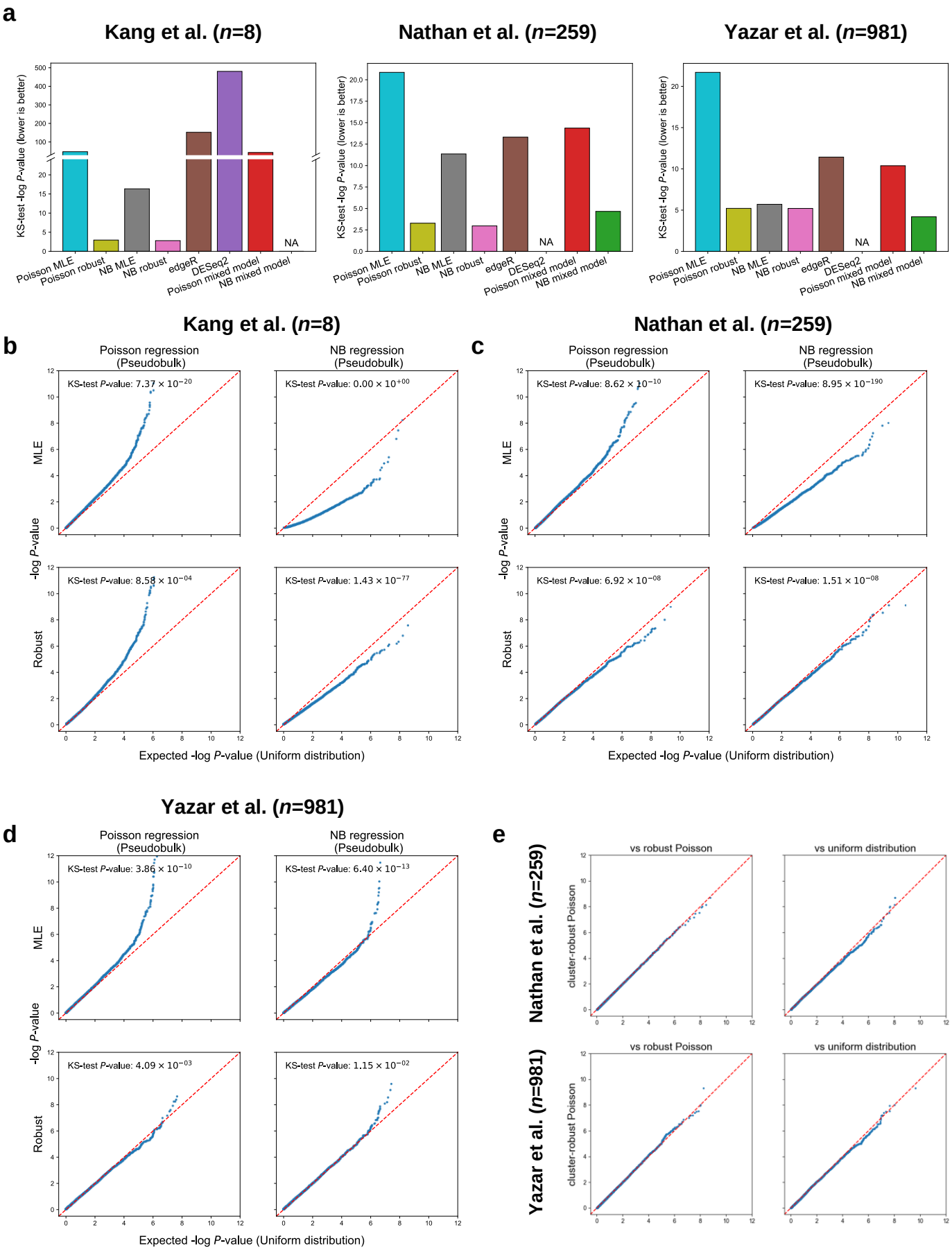
21 Bergé, L. Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. (Department of Economics at the University of Luxembourg, 2018).

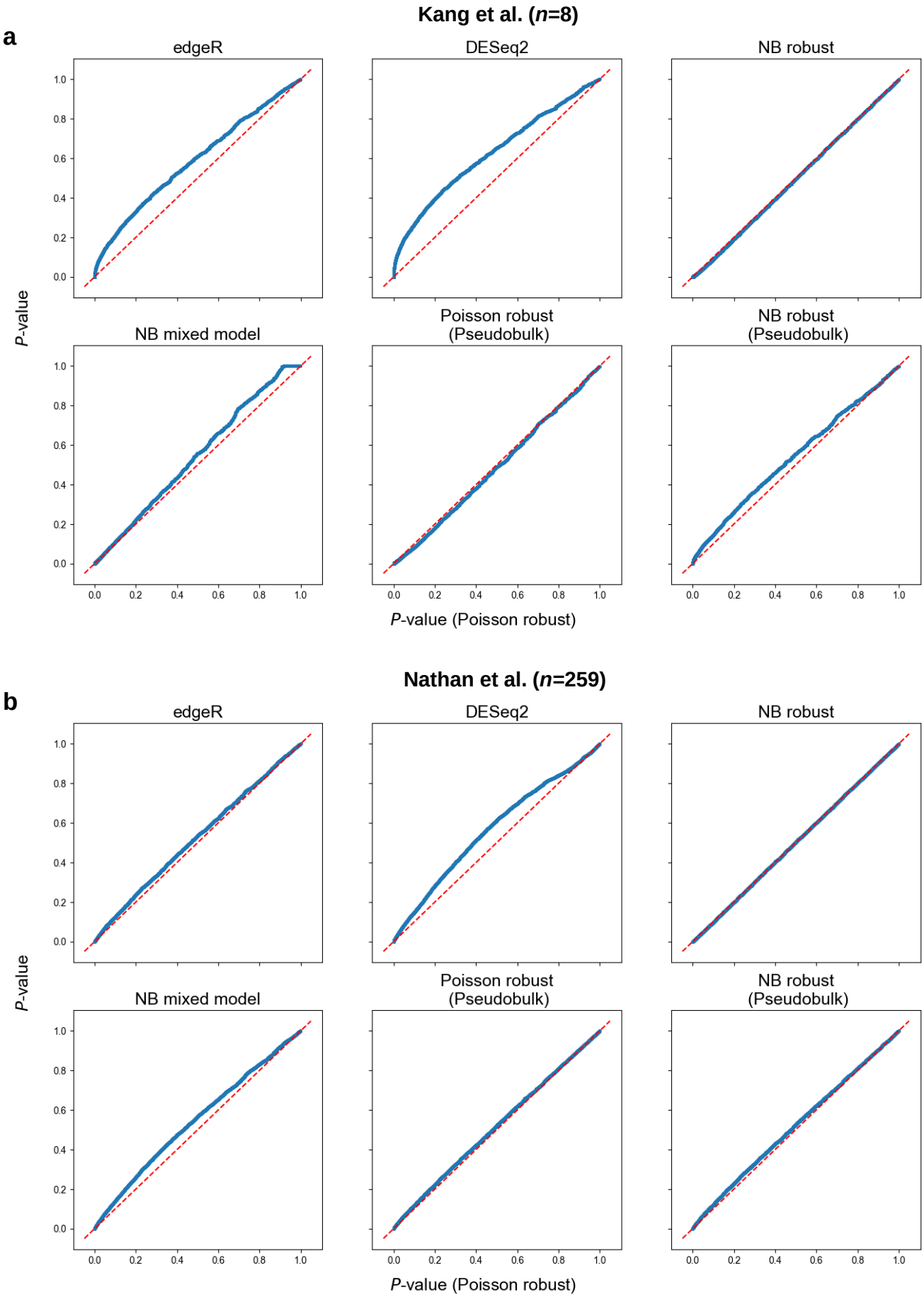
22 Wooldridge, J. M. in *Handbook of Applied Econometrics Volume 2: Microeconomics* 321-368 (1999).

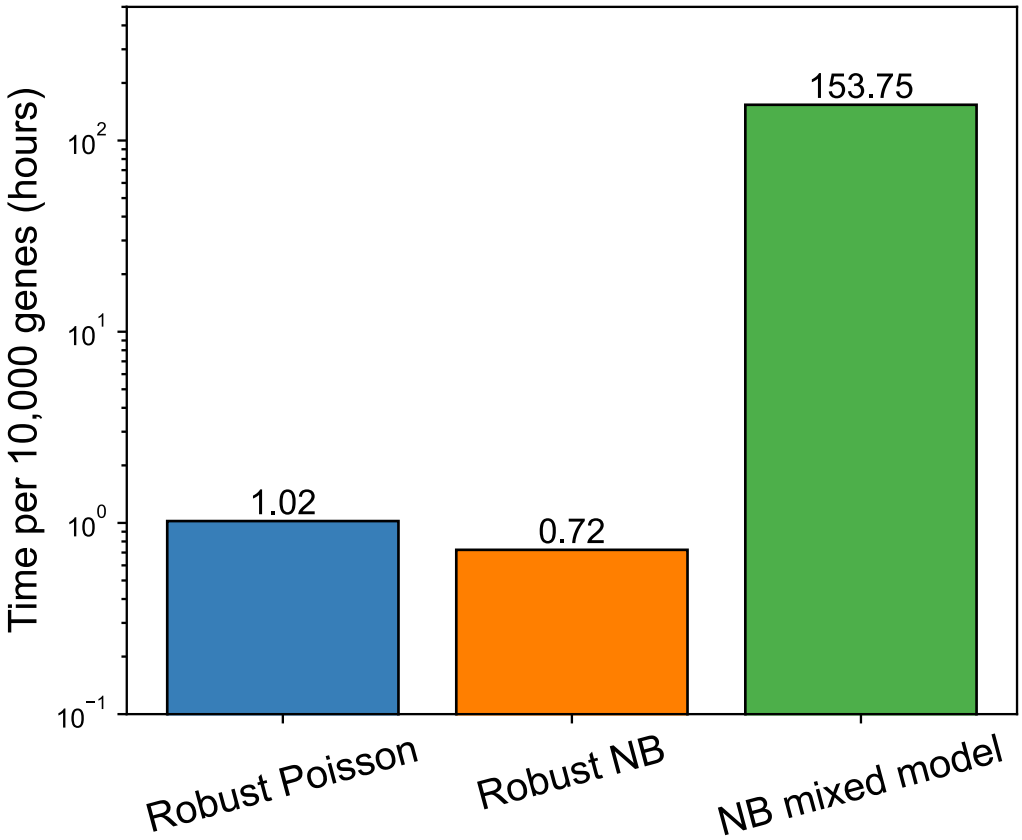
23 Lund, S. P., Nettleton, D., McCarthy, D. J. & Smyth, G. K. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat Appl Genet Mol Biol* **11**, doi:10.1515/1544-6115.1826 (2012).


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^−08
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^−02
NB robust
KS-test P-value: 6.08 × 10^−02
DESeq2
KS-test P-value: 1.91 × 10^−209
NB mixed model
Convergence failure
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)
NA
Poisson MLE
Poisson robust
NB MLE
NB robust
edgeR
DESeq2
Poisson mixed model
NB mixed model
b
Kang et al. (n=8)
Poisson regression (Pseudobulk)
NB regression (Pseudobulk)
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
c
Nathan et al. (n=259)
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
MLE
Robust
-log P-value
Expected -log P-value (Uniform distribution)
d
Yazar et al. (n=981)
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
e
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)


Time per 10,000 genes (hours)
10^{2}
10^{1}
10^{0}
10^{-1}
1.02
0.72
153.75
Robust Poisson
Robust NB
NB mixed model