Title: An appropriate differential expression analysis in single-cell data depends on the study design

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

Large-scale multi-subject single-cell data have become very common. However, these data have high heterogeneity in study designs, leading to confusion in differential gene expression (DGE) analysis. In this work, we show that pseudobulk produces a substantial amount of type 2 error when the group label being compared varies within a subject. When the comparison label is constant within a subject, type 1 error is inflated if pseudoreplication is not accounted for. As a general principle, we show that an appropriate method depends on the design of the study. We propose solutions to the inflated error rates and provide practical guidelines for researchers.

Introduction

Single-cell RNA sequencing (scRNA-seq) allowed the detection of cell-type specific differential expression patterns previously obscured in bulk RNA sequencing[1,2]. Now, multiple modalities are measured simultaneously in a single experiment[3]. Furthermore, massively parallel perturbation based on CRISPR technology is applied at a single-cell resolution[4]. Also, data are collected from both healthy and diseased individuals[5,6].

Nevertheless, these complexities raise a challenge in differential gene expression (DGE) analysis, which compares expression levels between prespecified groups of cells in different states[1,2]. Previous studies have warned of the danger of false discoveries due to the complicated data-generating process of scRNA-seq data[7,8]. One notable phenomenon is pseudoreplication bias, which emerges in multi-subject studies. Characteristics of cell donors (or experimental conditions in which the cells were obtained) can affect the expression level of cells, causing cells within a subject to be correlated. Pseudobulk and mixed models have been proposed as solutions based on simulations and empirical findings. However, it is still unclear under which conditions pseudoreplication is problematic and needs to be corrected[7,8].

In this work, we show that how the cell state is determined is a crucial factor for determining the appropriate DGE analysis strategy. There are two large categories of study designs in single-cell DGE analysis: the case-control design and the varying-cell-state design. In case-control designs where cells from healthy controls and diseased individuals are compared, the cell states of the cells are completely determined by the case-control status of the donors. On the contrary, the states of the cells vary within a donor in perturb-seq experiments. For example, a commonly used dataset of peripheral blood mononuclear cells (PBMC) for the DGE benchmark has varying cell states within a donor[9].

Pseudobulk remains reliable in the former design, where the state of their donor completely determines cell states. However, pseudobulk becomes overtly conservative in the latter design, where cell states are determined at a cell level. By contrast, cell-wise methods remain reliable in the latter design but become unreliable in the former design. Many studies in the literature chose the analysis strategies crossly, applying pseudobulk to the latter design or cell-wise methods to the former design, which can lead to uncalibrated type 1 and 2 errors. We found that mixed model remains valid in both designs as long as the sample size is sufficient but is often the slowest. We demonstrate these points through realistic simulation based on multi-subject datasets along with mathematical arguments.

Finally, we analyzed two publicly available experimental perturbation datasets[9,10]. When appropriately using mixed models and cell-level tests, we obtained more signals compared to using pseudobulk, which would be an inappropriate choice for this study design. As a concluding remark, we provide a simple guide for users willing to perform differential expression analysis in single-cell data.

Results

Overview

We argue that what is currently called single-cell DGE analysis is a heterogeneous mixture of completely different situations that should be dealt with differently. Broadly, single-cell DGE analysis can be divided into two large categories. In the first scenario, the comparison groups are assigned at the subject level. Case-control DGE analysis is an example in this category where cells from diseased and healthy individuals are compared. The second scenario assigns the group label at the cell level. In experimental perturbation studies, CRISPR or chemical perturbations are applied to individual cells, making variation of cell state within a subject[4,10,11].

The appropriate method for detecting differentially expressed genes (DEGs) depends ultimately on which category the study belongs to because the two scenarios have completely different data-generating processes. In scenario 1, a subject is sampled with a case/control label, and the cells are subsequently obtained from the sample. Hence, the comparison group (which is the case/control label) of cells is determined prior to the sampling of cells. By contrast, in scenario 2, the comparison group is assigned at the cell level after the subjects are obtained (e.g. CRISPR affected cells/unaffected cells).

These different data-generating processes naturally lead to different scales of standard errors of the effect size estimate (log fold change). Standard error is the square root of the estimator's variance, provided the data-generating process is repeated infinitely many times. Let n be the number of subjects and N be the total number of cells ($n \ll N$). In **Theorem 1 of Supplementary Note**, we show that under scenario 1, standard error is $O\left(\frac{1}{\sqrt{n}}\right)$, while under scenario 2, standard error is $O\left(\frac{1}{\sqrt{N}}\right)$. The smaller standard error in scenario 2 is intuitive, as the assignment of the comparison group is performed at a higher resolution (cell-wise).

In practice, we don't have access to the true variance of the estimator because the experiment is done only once. Instead, a DGE method estimates the variance using the data. The problem is that this estimate can be incorrect. With few exceptions like the mixed model, most statistical tests assume independence between the observations. As a result, pseudobulk always estimates the standard error as $O\left(\frac{1}{\sqrt{n}}\right)$ because the input data has n observations. Therefore, applying pseudobulk to scenario 2 leads to deflated type 1 errors and inflated type 2 errors. In contrast, cell-wise methods always approximate the standard error as $O\left(\frac{1}{\sqrt{N}}\right)$. Therefore, applying cell-wise methods to scenario 1 leads to inflated type 1 errors and deflated type 2 errors. This inflation of type 1 error was already reported elsewhere[2,7,8]. We confirmed these expected phenomena by both simulations and real data analysis.

There is a lot of confusion in single-cell DGE literature. Some studies applied pseudobulk to scenario 2[10,12,13], and some studies applied cell-wise methods to scenario 1[2,7,8,14]. Some studies simply describe the method name without any context[12]. For example, a widely used DGE method is glmGamPoi[15], and this can be applied both via pseudobulk and cell-wisely. Thus, stating that the study used glmGamPoi does not provide enough information on how the DGE test was conveyed.

Pseudobulk is frequently used in multi-subject scRNA-seq studies in the fear of pseudoreplication bias. Traditionally, pseudoreplication bias refers to the high false discovery rate of cell-wise methods applied to multi-subject studies[2,7,8]. We show that not all multi-subject studies are subject to the bias because the bias only happens in the first scenario (**Corollary 1 and 2 of Supplementary Note**). Subjects, not cells, are the true replicates in the first case, so pseudobulk is an effective DGE analysis method. Nevertheless, in the second case, cells are the true replicates, and pseudobulk is a suboptimal choice.

Our article offers theoretically grounded but practical guidance on which method is appropriate for which situation (**Figure 1**). Pseudobulk is appropriate for scenario 1 only. Cell-wise methods are suitable for scenario 2 only. Mixed model (NB GLMM) has an interesting position in that its type 1 and 2 errors are calibrated in both scenarios as long as the sample size is sufficient. However, NB GLMM is the slowest method of all methods compared and often fails to converge for low-expression genes. In the following sections, we present simulation results that support our guidelines.

Performance of single-cell DGE methods under simulations of two scenarios

We conducted null and power simulations to evaluate the performance of the methods. Simulations were conducted for both scenarios of comparison group assignment (subject-wise and cell-wise). We compared three pseudobulk methods (glmGamPoi with aggregated counts, edgeR, and limma), two cell-wise methods (robust GLM, glmGamPoi with individual cells), and one negative binomial mixed model (NB GLMM) (see **Methods**).

An ideal *P*-value follows a uniform distribution under the null hypothesis. This means that when identical experiments are repeated, the overall distribution of the resulting *P*-values should follow a uniform distribution. Quantile-quantile (QQ) plot is a convenient way to see if a collection of *P*-values follows a uniform distribution. In the plot, randomly drawn values from the uniform distribution and the test's *P*-values are placed in the *x*-*y* plane after ordering. The points are then found in the *y=x* line if the *P*-values follow the uniform distribution. Points above the line indicate that *P*-values are larger than expected, which means that the test is conservative. When the points are below the line, it means that the test is anti-conservative, leading to false positives. QQ-plot shows the distribution of *P*-values across all ranges (from 0 to 1), providing more comprehensive information than just reporting the type 1 error rate, which only summarizes the overall distribution of *P*-values at a particular significance threshold. Importantly, the significance threshold in DGE analysis usually varies across studies due to varying numbers of tests (e.g. number of transcripts being tested) and false discovery adjustment methods. Therefore, we presented our results in QQ-plots (left panels of **Figure 2 and 3**).

In the right panel, we presented the log-transformed *P*-values generated from the power simulation. In this case, the true difference between the comparison group is set to a non-zero value, so the *P*-values are smaller than those produced from the null simulation. The more powerful the test is, the smaller the produced *P*-values. To effectively visualize the small *P*-values, we applied -log to the numbers. A more powerful test, therefore, produces a larger -log *P*-value. Although higher power offers better sensitivity, it might result from sacrificing false discovery rate. Hence, a test that has higher power without controlling type 1 error properly should be interpreted with caution (e.g. cell-wise methods in scenario 1 in the following simulations).

Overall, pseudobulk methods and NB GLMM performed well in the first scenario (**Figure 2**, **Supplementary Figure 1 and 2**). In terms of type 1 error, pseudobulk methods gave calibrated error, except edgeR gave lower error than desired in several datasets. NB GLMM's *P*-values were deflated with small sample size (*n*=10), but they were improved with large sample size (*n*=50). In terms of power, NB GLMM was slightly more powerful than pseudobulk methods in some datasets. By contrast, cell-wise methods were severely anti-conservative, with many false positives in the null simulation. Although cell-wise methods look more powerful according to the *P*-values in the right panel of **Figure 2**, this should be interpreted along with the failure to control type 1 errors.

In the second scenario where the cell states vary within a subject, cell-wise methods and NB GLMM exhibited well-calibrated *P*-values under the null (**Figure 3**, **Supplementary Figure 3 and 4**). Pseudobulk methods, on the other hand, were generally too conservative and often produced overly large *P*-values under the null, significantly diverging from the expected distribution. In the power simulation, cell-wise methods and NB GLMM's *P*-values were much smaller than those of pseudobulk, demonstrating these should be the methods of choice for this scenario.

Analysis of publicly available experimental perturbation datasets

A variety of single-cell perturbation assays have been developed. As shown in the previous section, these data (scenario 2) have a different data-generating process, so DGE analysis should be performed differently from case-control single-cell studies (scenario 1). If not, the chance of missing true signal can increase substantially.

To demonstrate this point, we analyzed two perturbation datasets using the 6 methods compared in the simulations (see **Methods**). Note that these datasets have been frequently analyzed through pseudobulk approaches[10,12,16-18]. Under a stringent significance threshold (p<0.05/6257), pseudobulk methods rejected less than 7% transcripts, while cell-wise methods and NB GLMM rejected more than 11.5% in an interferon-g (IFN-g) stimulation experiment (**Figure 4a**). Similarly, in a CRISPR perturbation experiment, pseudobulk methods failed to find any signals, while cell-wise methods and NB GLMM found more than 10% of the genes to be differentially expressed (p<0.05/4807, **Figure 4b**).

Nevertheless, the rank of genes ordered by their *P*-values was consistent to some degree. As can be expected, the concordance was higher within pseudobulk methods and within cell-wise (or NB GLMM) methods (**Figure 4a** and **Figure 4b**), reaching pairwise correlations higher than 0.9 within these groups. Across these groups, the Spearman rank correlations dropped down to 0.83. Although 0.83 looks high, this imperfect (<1) correlation implies that the ranks of key genes can change.

Discussion

We demonstrated the importance of study design in single-cell DGE analysis. The seemingly common data structure in single-cell studies, where comparison groups are discrete random variables (e.g. dummy variables in 0 and 1) and the outcome variables are the observed transcript counts, has confused researchers. The study design, the mechanism of how these discrete group variables are assigned to the cells, has a profound consequence on DGE analysis. When cell states are assigned at the subject level, cells within a subject can be treated as a single observation as a whole via pseudobulk. On the contrary, they can be treated as independent observations when each cell is assigned its state independently (**Supplementary Note**).

The ramification of ignoring study design comes in both ways. False discoveries are severely inflated when cells are treated as independent observations when they aren't, and true DEGs are missed when independent observations are treated as aggregates. For example, pseudobulk is effective in accounting pseudoreplication bias in case-control studies but is inadequate for perturb-seq experiments. Cell-level methods are powerful and valid for perturb-seq experiments but are fragile when applied to case-control studies.

We proposed a theoretical argument on the origin of pseudoreplication bias (**Supplementary Note**). Most, if not all, statistical estimators' variance is inversely proportional to the number of independent replicates due to the central limit theorem[19]. We've shown that this number is the number of samples (n) and the number of cells (N) in the first and the second scenarios, respectively. In case-control designs, therefore, the subjects are the true replicates, and the cells are fake ones, which explains the name of the pseudoreplication bias. In the second design, the estimator converges to the true parameter at a $\sqrt{n}$-rate, so cells are true replicates in this case. Therefore, cell-wise methods correctly estimated the standard errors. Also, pseudoreplication bias does not occur in the second case.

Negative-binomial generalized linear mixed model (NB GLMM) was the only calibrated method in both designs, given a sufficient sample size. Although the high computational demand often makes it less attractive, a scalable implementation called NEBULA has been introduced[20]. One weakness is that NB GLMM often fails to converge. We found that this happens frequently in transcripts with low mean counts (<0.1) (**Supplementary Figure 5**). If NB GLMM fails to converge, replacing it with other methods for the low-expressed transcripts is a putative solution.

scRNA-seq data are generally sparse, i.e., they have many zeros[21]. Dropout has been considered a potential cause, and several methods were developed to address this phenomenon in DGE analysis[22,23]. We investigated the performance of two methods (MAST mixed model and DEsingle). In our simulations, DEsingle was anti-conservative and conservative in scenarios 1 and 2, respectively. This was expected as DEsingle infers the number of replicates directly from the dimension of the input matrix. MAST performed well when the number of subjects was small but was often conservative when the subject number grew (**Supplementary Figure 6 and 7**). Also, severe inflation of false discoveries was observed in the second scenario. More recent works suggest that the combination of the discreteness and the low mean of the count distribution causes the observed sparsity[21,24-26]. Also, zeros from the low count distribution and the dropout component may not be distinguishable solely through statistical procedures[21,24]. The latest normalization method SCTransform of Seurat also does not consider zero-inflation in UMI datasets[27]. Therefore, we raise caution on using zero-inflated models.

The cell-wise and pseudobulk methods evaluated in the paper are all distribution-robust methods. edgeR and glmGamPoi implement a quasi-likelihood ratio test [15,28,29]. By quasi, it means that the likelihood does not require the data to exactly follow the assumed distribution (i.e. negative binomial or Poisson). Robust GLM implements a robust z-test (or equivalently, the Wald test), which has the same robustness to distributional misspecification[30-32]. limma fits a linear model, but non-Gaussian data is handled similarly, making it robust against distributional violations[33]. As single-cell data consists of diverse cell populations, a simple parametric probabilistic model may not suffice. This makes robust methods an attractive choice. Additionally, in the **Supplementary Note**, we provide a theoretical argument as to why non-dropout models still work when dropouts do occur. In **Supplementary Figure 8-10**, QQ-plots of vanilla Poisson regression and negative-binomial regression under various violations of the distributional assumptions can be found. Robust GLM is the only method that remains calibrated in all scenarios.

We expect the importance of study design to grow in future studies as the complexity of single-cell studies increases. The validity of the methods should be assessed for each design separately to prevent type 1 and 2 errors.

Methods

Datasets in null and power simulations

Three datasets were used: Reichart et al. (human heart)[34], Yazar et al. (human blood)[35], and Reed et al. (human breast)[36]. We used the cell type label provided by the author. We began with the datasets provided by the authors. For each dataset, we selected the top 6 cell types and retained donors with more than 50 cells for each cell type. This left 76, 165, and 47 donors in Reichart et al., Yazar et al., and Reed et al., respectively. No further removal of data was made. The following table contains the link to the datasets.

Dataset	Link
Reichart et al.	https://cellxgene.cziscience.com/collections/e75342a8-0f3b-4ec5-8ee1-245a23e0f7cb
Yazar et al.	https://cellxgene.cziscience.com/collections/dde06e0f-ab3b-46be-96a2-a8082383c4a1
Reed et al.	https://cellxgene.cziscience.com/collections/48259aa8-f168-4bf5-b797-af8e88da6637

The codes used in the analysis can be found at https://github.com/hanbin973/DEGpaper.

Null simulations

In multi-subject scRNA-seq studies, cell donors are first selected from the population. Next, cells are sampled from each selected individual. In our simulation, we mimic this process by sampling individuals from the data and subsequently selecting cells from the sampled individuals. The individuals were selected with replacement to prevent artificial negative correlation between cells across individuals when the number of donors is finite[37].

In the case of constant cell states within a donor, the first simulation assigned all cells from the same donor to the same state. Note that the state was defined as a binary variable (used R's rbinom function). Each sampled individual was randomly allocated to either group with equal probability (p=0.5). In the second scenario, where cell state varies within a donor, cell states were assigned for each cell independently with equal probability (p=0.5, using rbinom function).

The methods evaluated in the main figure were negative binomial generalized linear mixed model (NB GLMM), glmGamPoi (with and without pseudobulk), edgeR (pseudobulk), limma (pseudobulk), and robust GLM. For NB GLMM and robust GLM, we used NEBULA and fixest software[15,20,29,33,38], respectively. All methods were run with their default setting except for fixest (vcov='robust' option in fepois function). All pseudobulk was created by summing up the counts within a replicate using the pseudobulk function in glmGamPoi library. Except for limma, all methods took raw UMI counts as inputs. edgeR internally performs normalization with the raw UMI counts and voom normalization was done for limma using the \`voom\` function of the library.

10 iterations were performed for 100 randomly selected genes in each cell-type. To prevent convergence failures in mixed models, we excluded genes with a mean smaller than 0.1 for each cell type. **Supplementary Figure 5** reports the frequency of convergence failure of NB GLMM in the Kang et al. dataset. For completeness, two implementations were tested (NEBULA and glmmTMB[39]).

Power simulations

Power simulation was performed identically to the null simulation except for the following step. After the treatment was assigned, we randomly selected 5 genes and downsampled their count according to the prespecified fold-change (per iteration). The downsampling formula is

$$\text{New count} \sim \text{BinomialDistribution}(n = \text{Original count},\ p = \text{fold change})$$

The fold change was set to 50% in all simulations.

Note that multiple genes were analyzed simultaneously per iteration to reflect the effect normalization. Methods frequently perform normalization using the total count across all transcripts present in the dataset. Earlier studies have ignored this step by testing one gene at a time [7,14]. When a zero *P*-value occurred, we rounded them to 10^{-300} to produce the plots.

In both null and power simulations, the average number of cells of the least frequent cell types were 7481, 971 and 7511 in Reichart et al., Yazar et al., and Reed et al. datasets, respectively.

Perturbation datasets

We adopted the code from Schmidt et al.[10]. All cells from the data were used without further filtering. Cells were divided into separate cell-types and genes with mean larger than 0.1 were selected. The links to the data and codes are listed in the following table.

Dataset	Link
Kang et al.	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE96583
Schmidt et al.	https://zenodo.org/records/5784651

NB GLMM, glmGamPoi (with and without pseudobulk), edgeR (pseudobulk), limma (pseudobulk), and robust GLM (without pseudobulk) were compared. In Schmidt et al., pseudobulk was defined using the lane variable in the metadata. In Kang et al., the donor variable was used to define pseudobulk.

Evaluation of methods considering dropouts

MAST and DEsingle were evaluated[22,23,40]. The procedure was identical to other methods in the null and the power simulations. Log normalization of the data before applying MAST was done using the scuttle library (logNormCounts function)[22,23,40]. The comparison was performed in the Yazar et al. dataset.

Robust GLM and negative-binomial quasi-likelihood methods

Robust GLM of fixest[30] and negative-binomial (NB) quasi-likelihood ratio test (QLRT) of edgeR and glmGamPoi[28] are robust to distributional misspecifications. We provide a brief description of these methods. Poisson regression and negative-binomial regression are commonly implemented through maximum likelihood estimation (MLE). To perform MLE, the precise probability density (or mass) function should be known. If not, the procedure may produce bias. Although not exhaustive, bias in the point estimate (logFC in DGE) and the variance estimate of the estimator (the square of standard error) are two notable examples.

Poisson regression's MLE estimate is correct, but the standard errors are generally deflated in overdispersed data. NB regression's MLE estimate may suffer from both biases at the same time if data is not negative-binomial distributed. Therefore, both robust GLM and NB QLRT retain the point estimate of Poisson MLE and correct the standard error to carry out inference when the distribution is not exactly known. The correction methods are slightly different in the two cases, but the resulting *P*-values were highly concordant in our simulations (**Figure 2-3 and Supplementary Figure 1-4**). We highly recommend Wooldridge (1999)[41] for the overview of various robust tests for count data. Lund et al. (2012)[28] contains the details on the Bayesian shrinkage method for overdispersion estimates of NB QLRT.

Except for the name, NB QLRT and the usual NB regression are very different. NB regression estimates the regression coefficient and the overdispersion parameter simultaneously. For each iteration in optimizing the NB log-likelihood function, two parameters are optimized simultaneously. On the other hand, NB QLRT first estimates the regression coefficient using the Poisson log-likelihood function. Overdispersion is then estimated after fixing the regression coefficient obtained from the previous step[28,41,42]. The only common feature of the two methods giving NB QLRT its name is the same mean-variance relationship used to carry out inference, usually called the NB2 parameterization[43].

Figure legend

Figure 1. Recommendation of methods according to the study design.

Figure 2. Null and power simulation when comparison group is assigned at the subject level (scenario 1) in the Reed et al. dataset. **a.** Observed versus expected *P*-values in the null simulation. **b.** Observed versus uniform -log*P* values.

Figure 3. Null and power simulation when comparison group is assigned at the cell level (scenario 2) in the Reed et al. dataset. **a.** Observed versus expected *P*-values in the null simulation. **b.** Observed versus uniform -log*P* values.

Figure 4. Heatmap of proportion (%) of null rejections and the Spearman correlation between methods. **a.** IFN-g stimulation of peripheral blood mononuclear cells (PBMC)[9]. **b.** CRISPR perturbation of T-cells[10]. In the left panel, the diagonal values are the proportion of rejected nulls of the methods. The non-diagonal values are the proportion of nulls that were rejected by both methods. The values of the right panel are pairwise the Spearman correlation of *P*-values.

Supplementary Figure legend

Supplementary Figure 1. Null and power simulation when comparison group is assigned at the subject level (scenario 1) in the Yazar et al. dataset. **a.** Observed versus expected *P*-values in the null simulation. **b.** Observed versus uniform -log*P* values.

Supplementary Figure 2. Null and power simulation when comparison group is assigned at the subject level (scenario 1) in the Reichart et al. dataset. **a.** Observed versus expected *P*-values in the null simulation. **b.** Observed versus uniform -log*P* values.

Supplementary Figure 3. Null and power simulation when comparison group is assigned at the cell level (scenario 2) in the Yazar et al. dataset. **a.** Observed versus expected *P*-values in the null simulation. **b.** Observed versus uniform -log*P* values.

Supplementary Figure 4. Null and power simulation when comparison group is assigned at the cell level (scenario 2) in the Reichart et al. dataset. **a.** Observed versus expected *P*-values in the null simulation. **b.** Observed versus uniform -log*P* values.

Supplementary Figure 5. Convergence failure rate versus transcript mean of negative-binomial mixed model. **a.** glmmTMB implementation. **b.** NEBULA implementation.

Supplementary Figure 6. Null and power simulation when comparison group is assigned at the subject level (scenario 1) in the Yazar et al. dataset. The methods are MAST mixed-model and DEsingle. **a.** Observed versus expected *P*-values in the null simulation. **b.** Observed versus uniform -log*P* values.

Supplementary Figure 7. Null and power simulation when comparison group is assigned at the cell level (scenario 2) in the Yazar et al. dataset. The methods are MAST mixed-model and DEsingle. **a.** Observed versus expected *P*-values in the null simulation. **b.** Observed versus uniform -log*P* values.

Supplementary Figure 8. QQ-plots of Poisson MLE, robust GLM, and negative binomial (NB) MLE under negative-binomial distribution with varying dispersion parameter θ =1,2,5 in $Var(y) = \mu + \mu^2/\theta$ (NB2 parameterization).

Supplementary Figure 9. QQ-plots of Poisson MLE, robust GLM, and negative binomial (NB) MLE under zero-inflated Poisson distribution with varying zero inflation rate π=0.1, 0.2, 0.5.

Supplementary Figure 10. QQ-plots of Poisson MLE, robust GLM, and negative binomial (NB) MLE under a mixture of Poisson distributions with varying numbers of mixture components *n*=2,4,6.

Reference

1 Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**, e8746, doi:10.15252/msb.20188746 (2019).

2 Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nat Rev Genet* **24**, 550-572, doi:10.1038/s41576-023-00586-w (2023).

3 Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**, 865-868, doi:10.1038/nmeth.4380 (2017).

4 Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866 e1817, doi:10.1016/j.cell.2016.11.038 (2016).

5 Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332-337, doi:10.1038/s41586-019-1195-2 (2019).

6 Mathys, H. *et al.* Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. *Cell* **186**, 4365-4385 e4327, doi:10.1016/j.cell.2023.08.039 (2023).

7 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).

8 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).

9 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).

10 Schmidt, R. *et al.* CRISPR activation and interference screens decode stimulation responses in primary human T cells. *Science* **375**, eabj4008, doi:10.1126/science.abj4008 (2022).

11 Srivatsan, S. R. *et al.* Massively multiplex chemical transcriptomics at single-cell resolution. *Science* **367**, 45-51, doi:10.1126/science.aax6234 (2020).

12 Junttila, S., Smolander, J. & Elo, L. L. Benchmarking methods for detecting differential states between conditions from multi-subject single-cell RNA-seq data. *Brief Bioinform* **23**, doi:10.1093/bib/bbac286 (2022).

13 Santinha, A. J. *et al.* Transcriptional linkage analysis with in vivo AAV-Perturb-seq. *Nature* **622**, 367-375, doi:10.1038/s41586-023-06570-y (2023).

14 Murphy, A. E. & Skene, N. G. A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. *Nat Commun* **13**, 7851, doi:10.1038/s41467-022-35519-4 (2022).

15 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).

16 Crowell, H. L. *et al.* muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat Commun* **11**, 6077,

doi:10.1038/s41467-020-19894-4 (2020).

17 Helena L Crowell, C. S., Pierre-Luc Germain, and Mark D Robinson. *Differential state analysis with muscat*, <https://www.bioconductor.org/packages/devel/bioc/vignettes/muscat/inst/doc/analysis.html> (2023).

18 Ahlmann-Eltze, C. *glmGamPoi Quickstart*, <https://bioconductor.org/packages/devel/bioc/vignettes/glmGamPoi/inst/doc/glmGamPoi.html> (2023).

19 Durrett, R. *Probability: Theory and Examples* 4th Edition edn, (Cambridge University Press, 2010).

20 He, L. *et al.* NEBULA is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data. *Commun Biol* **4**, 629, doi:10.1038/s42003-021-02146-6 (2021).

21 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).

22 Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16**, 278, doi:10.1186/s13059-015-0844-5 (2015).

23 Miao, Z., Deng, K., Wang, X. & Zhang, X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics* **34**, 3223-3224, doi:10.1093/bioinformatics/bty332 (2018).

24 Choi, K., Chen, Y., Skelly, D. A. & Churchill, G. A. Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. *Genome Biol* **21**, 183, doi:10.1186/s13059-020-02103-2 (2020).

25 Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nat Biotechnol* **38**, 147-150, doi:10.1038/s41587-019-0379-5 (2020).

26 Jiang, R., Sun, T., Song, D. & Li, J. J. Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biol* **23**, 31, doi:10.1186/s13059-022-02601-5 (2022).

27 Choudhary, S. & Satija, R. Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biol* **23**, 27, doi:10.1186/s13059-021-02584-9 (2022).

28 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).

29 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).

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

31 Pesaran, M. H. & Schmidt, P. *Handbook of Applied Econometrics Volume II: Microeconomics*. (1999).

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

33 Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29, doi:10.1186/gb-2014-15-2-r29 (2014).

34 Reichart, D. *et al.* Pathogenic variants damage cell composition and single cell transcription in cardiomyopathies. *Science* **377**, eabo1984, doi:10.1126/science.abo1984 (2022).

35 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).

36 Reed, A. D. *et al.* A Human Breast Cell Atlas Mapping the Homeostatic Cellular Shifts in the Adult Breast. *bioRxiv*, doi:10.1101/2023.04.21.537845 (2023).

37 Fuller, W. A. *Sampling Statistics*. 1st edn, (Wiley, 2009).

38 Bergé, L. Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. (2018).

39 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).

40 McCarthy, D. J., Campbell, K. R., Lun, A. T. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179-1186, doi:10.1093/bioinformatics/btw777 (2017).

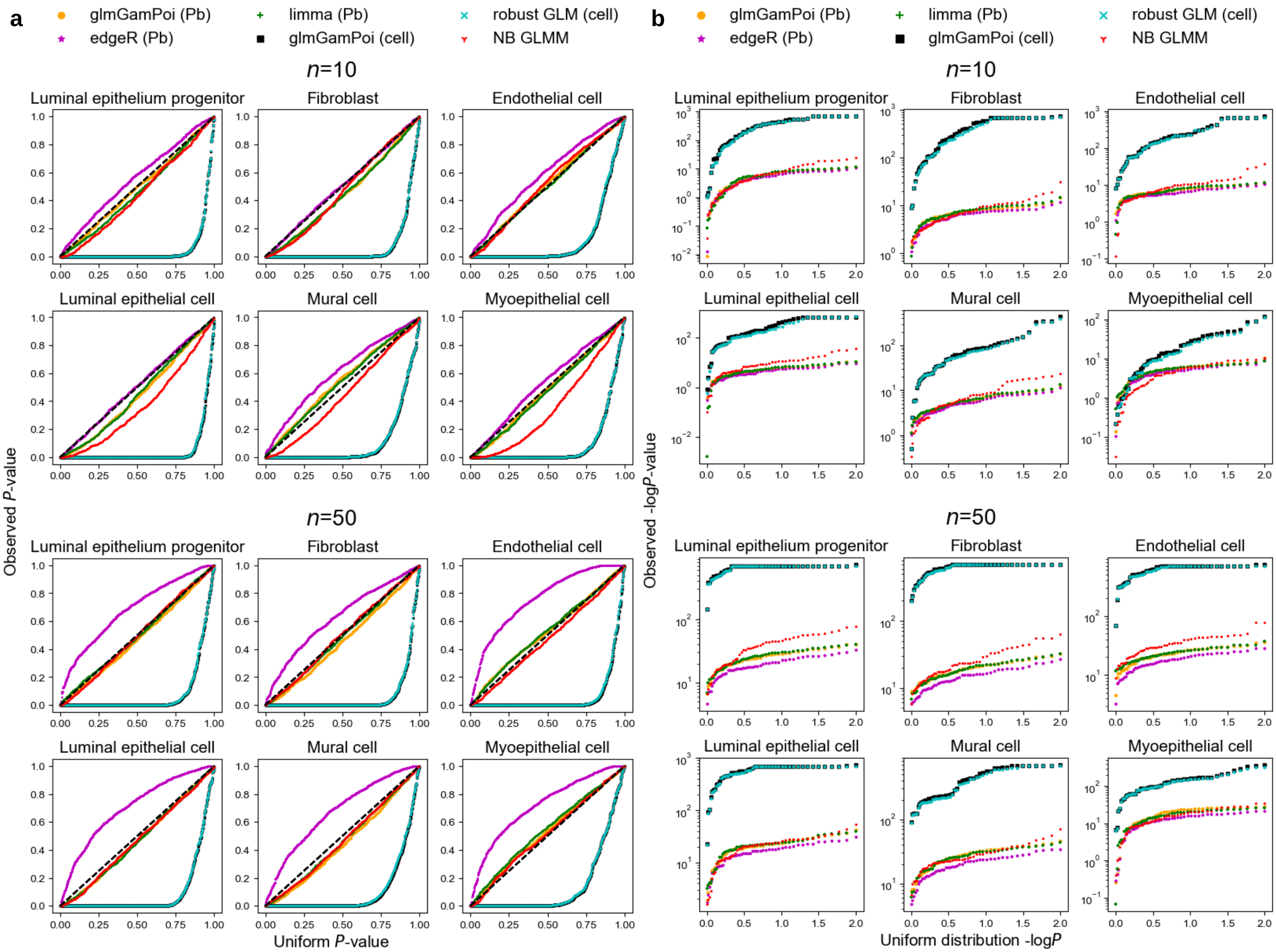
41 Wooldridge, J. M. in *Handbook of Applied Econometrics Volume 2: Microeconomics* (ed Peter Schmidt M. Hashem Pesaran) (John Wiley & Sons, 1999).

42 Ahlmann-Eltze, C. *Package 'glmGamPoi'*, <https://bioconductor.org/packages/release/bioc/manuals/glmGamPoi/man/glmGamPoi.pdf> (2023).

43 Greene, W. Functional forms for the negative binomial model for count data. *Economics Letters* **99**, 585-590, doi:10.1016/j.econlet.2007.10.015 (2008).

Figure 1

Method	Internal test	Comparison group assigned at subject level (Scenario 1)		Comparison group assigned at cell level (Scenario 2)		Runtime	Appropriate for Scenario 1	Appropriate for Scenario 2
		Type 1 error	Power	Type 1 error	Power			
glmGamPoi (Pseudobulk)	Quasi-likelihood ratio test	Calibrated	High	**Deflated**	**Deflated**	**Fast**	O	X
limma (Pseudobulk)	Least Square	Calibrated	High	**Deflated**	**Deflated**	**Fast**	O	X
edgeR (Pseudobulk)	Quasi-likelihood ratio test	Low	Low	**Deflated**	**Deflated**	**Fast**	O	X
NEBULA (Mixed model)	Negative binomial mixed model	Calibrated (when *n* large)	High	Calibrated	High	**Slow**	O (when *n* large)	O
glmGamPoi (cell-wise)	Quasi-likelihood ratio test	**Inflated**	**Inflated**	Calibrated	High	Medium	X	O
robust GLM (cell-wise)	Robust Wald test	**Inflated**	**Inflated**	Calibrated	High	Medium	X	O


a
b
● glmGamPoi (Pb)
★ edgeR (Pb)
+ limma (Pb)
■ glmGamPoi (cell)
× robust GLM (cell)
Y NB GLMM
n=10
n=50
Luminal epithelium progenitor
Fibroblast
Endothelial cell
Luminal epithelial cell
Mural cell
Myoepithelial cell
Observed P-value
Uniform P-value
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.25
0.50
0.75
1.00
Observed -logP-value
Uniform distribution -logP
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
0.0
0.5
1.0
1.5
2.0

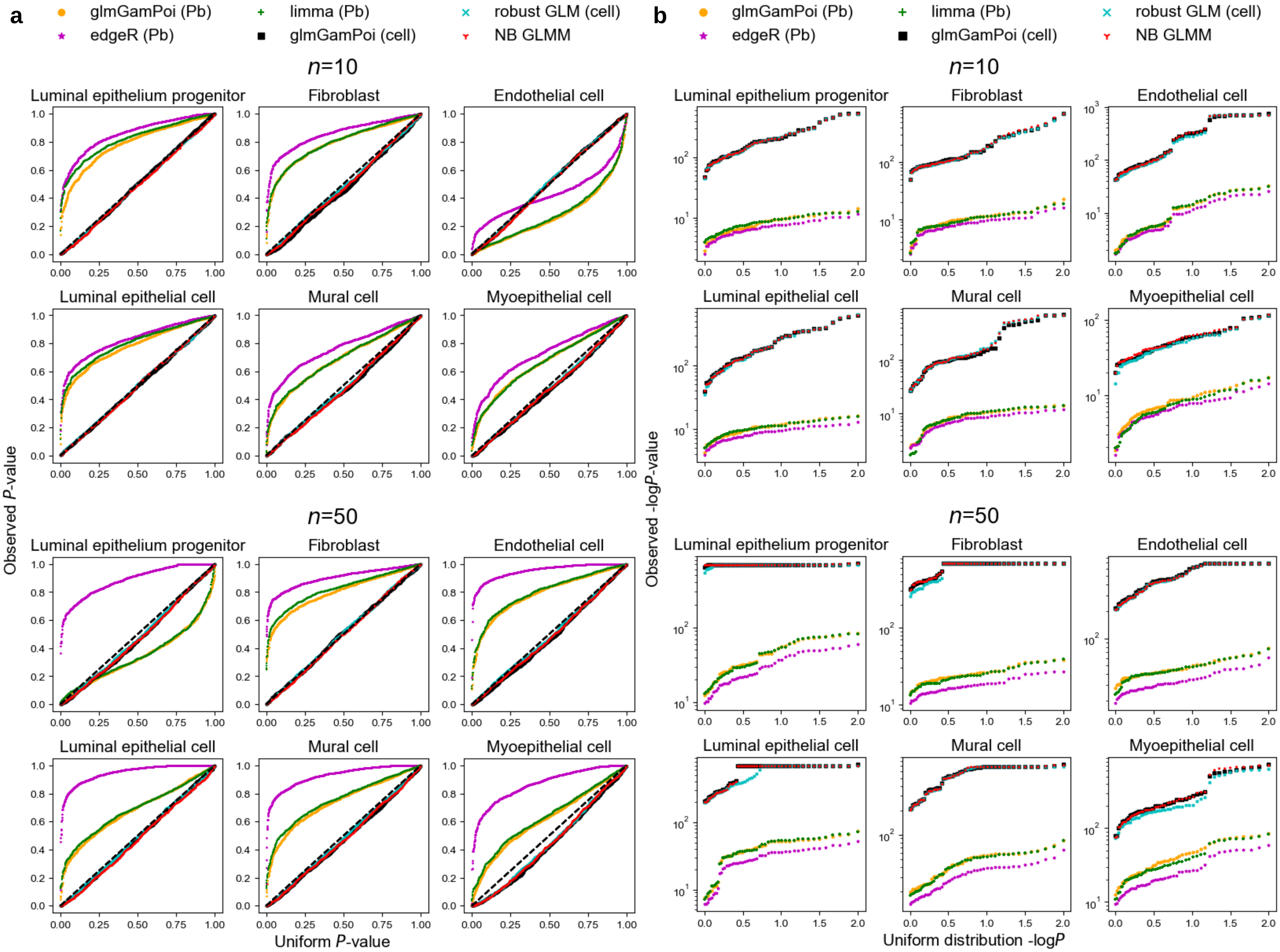
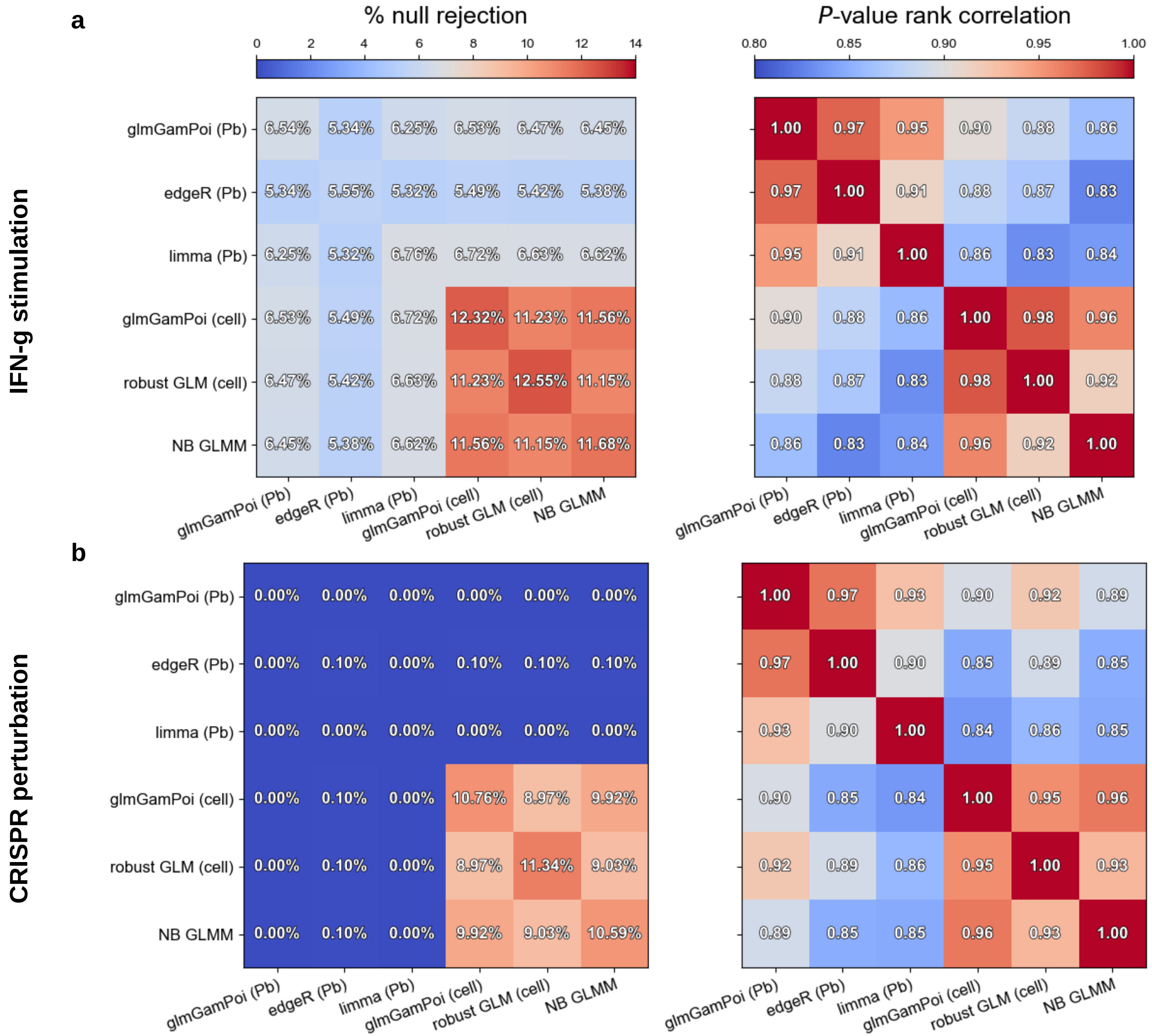
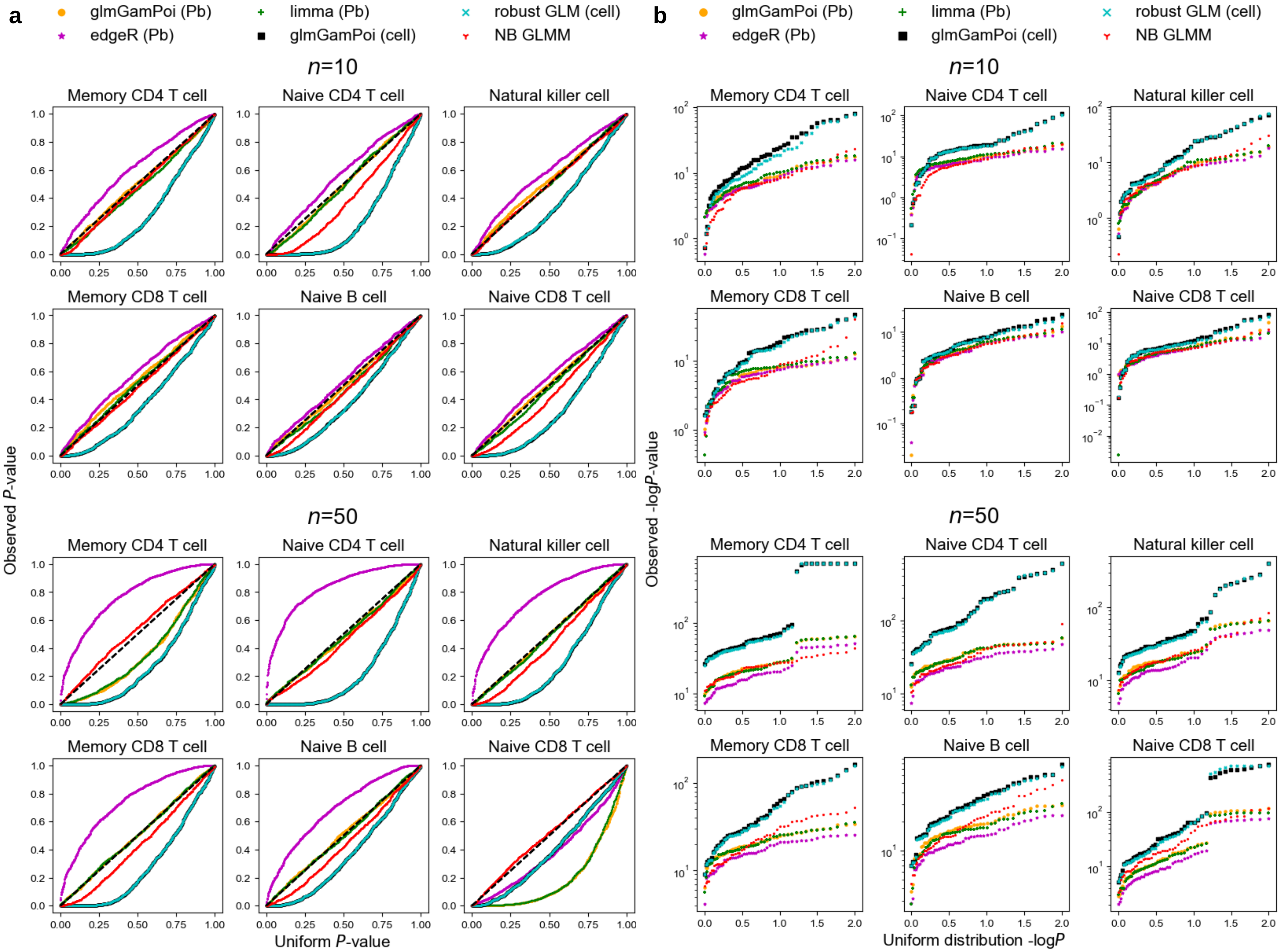
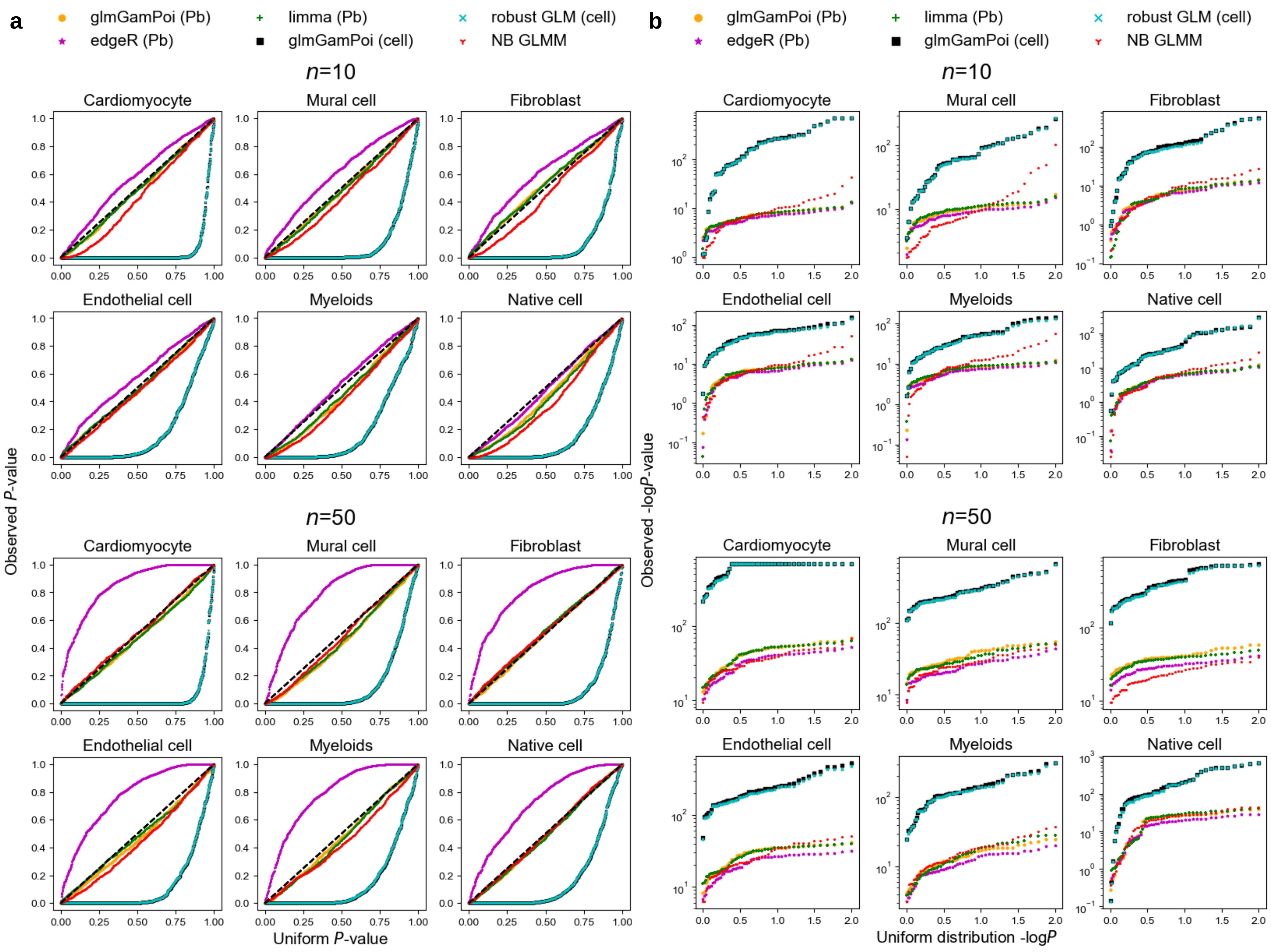

a
glmGamPoi (Pb)
edgeR (Pb)
limma (Pb)
glmGamPoi (cell)
robust GLM (cell)
NB GLMM
n=10
Luminal epithelium progenitor
Fibroblast
Endothelial cell
Luminal epithelial cell
Mural cell
Myoepithelial cell
n=50
Luminal epithelium progenitor
Fibroblast
Endothelial cell
Luminal epithelial cell
Mural cell
Myoepithelial cell
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.25
0.50
0.75
1.00
Observed P-value
Uniform P-value
b
glmGamPoi (Pb)
edgeR (Pb)
limma (Pb)
glmGamPoi (cell)
robust GLM (cell)
NB GLMM
n=10
Luminal epithelium progenitor
Fibroblast
Endothelial cell
Luminal epithelial cell
Mural cell
Myoepithelial cell
n=50
Luminal epithelium progenitor
Fibroblast
Endothelial cell
Luminal epithelial cell
Mural cell
Myoepithelial cell
10^1
10^2
10^3
0.0
0.5
1.0
1.5
2.0
Observed -logP-value
Uniform distribution -logP

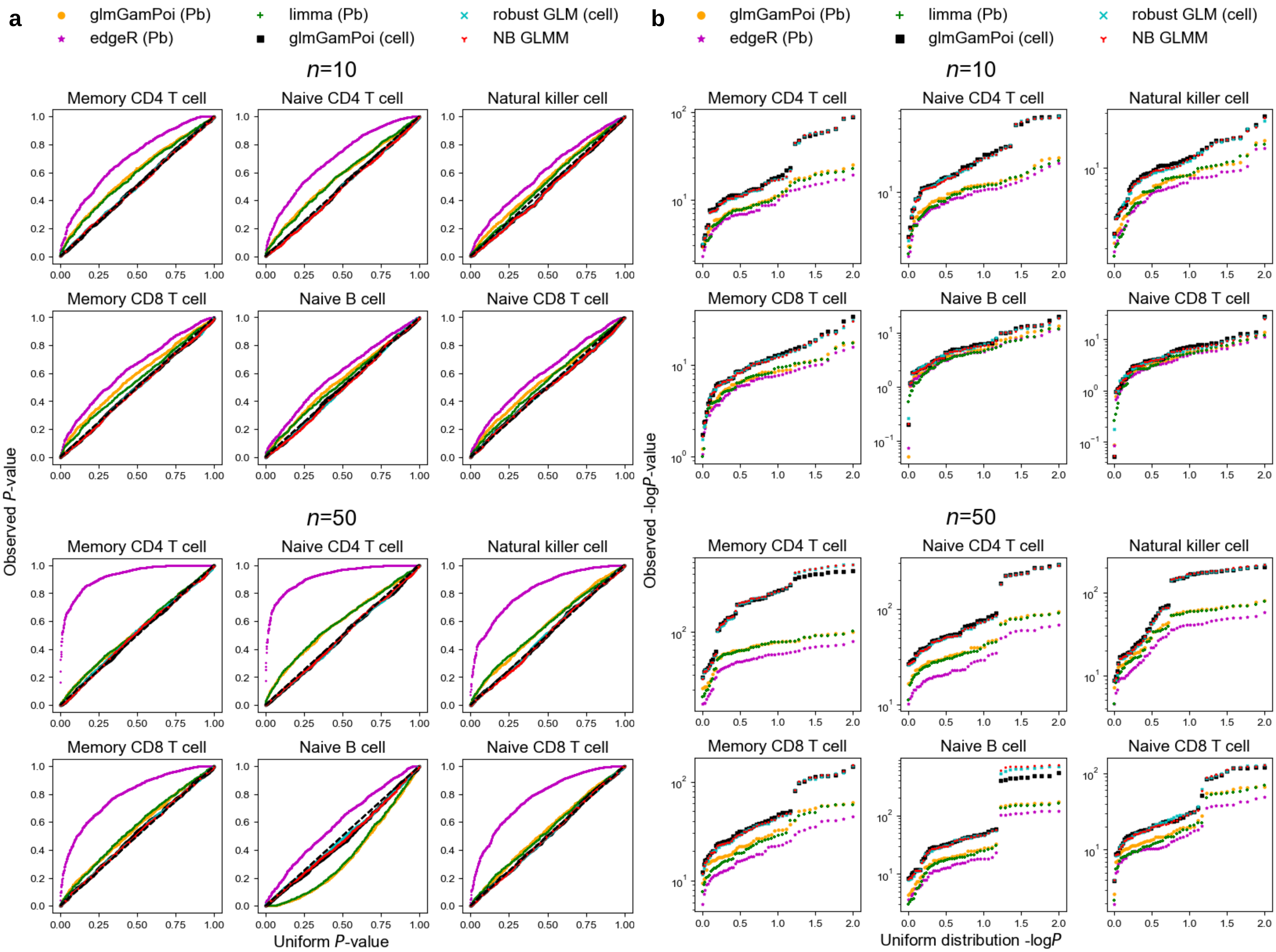
Figure 4

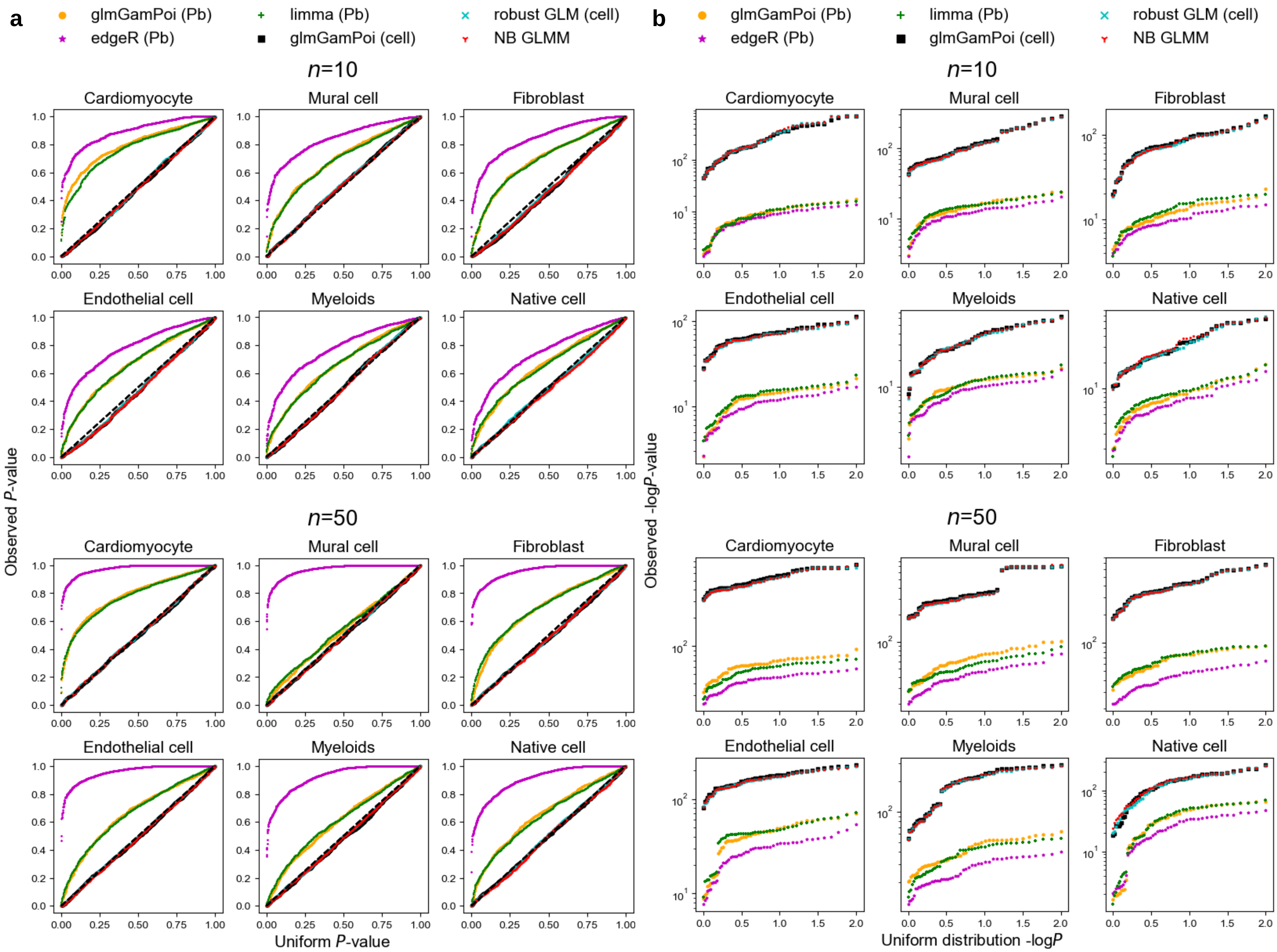

a
IFN-g stimulation
% null rejection
0 2 4 6 8 10 12 14
glmGamPoi (Pb) 6.54% 5.34% 6.25% 6.53% 6.47% 6.45%
edgeR (Pb) 5.34% 5.55% 5.32% 5.49% 5.42% 5.38%
limma (Pb) 6.25% 5.32% 6.76% 6.72% 6.63% 6.62%
glmGamPoi (cell) 6.53% 5.49% 6.72% 12.32% 11.23% 11.56%
robust GLM (cell) 6.47% 5.42% 6.63% 11.23% 12.55% 11.15%
NB GLMM 6.45% 5.38% 6.62% 11.56% 11.15% 11.68%
glmGamPoi (Pb) edgeR (Pb) limma (Pb) glmGamPoi (cell) robust GLM (cell) NB GLMM
P-value rank correlation
0.80 0.85 0.90 0.95 1.00
1.00 0.97 0.95 0.90 0.88 0.86
0.97 1.00 0.91 0.88 0.87 0.83
0.95 0.91 1.00 0.86 0.83 0.84
0.90 0.88 0.86 1.00 0.98 0.96
0.88 0.87 0.83 0.98 1.00 0.92
0.86 0.83 0.84 0.96 0.92 1.00
glmGamPoi (Pb) edgeR (Pb) limma (Pb) glmGamPoi (cell) robust GLM (cell) NB GLMM
b
CRISPR perturbation
glmGamPoi (Pb) 0.00% 0.00% 0.00% 0.00% 0.00% 0.00%
edgeR (Pb) 0.00% 0.10% 0.00% 0.10% 0.10% 0.10%
limma (Pb) 0.00% 0.00% 0.00% 0.00% 0.00% 0.00%
glmGamPoi (cell) 0.00% 0.10% 0.00% 10.76% 8.97% 9.92%
robust GLM (cell) 0.00% 0.10% 0.00% 8.97% 11.34% 9.03%
NB GLMM 0.00% 0.10% 0.00% 9.92% 9.03% 10.59%
glmGamPoi (Pb) edgeR (Pb) limma (Pb) glmGamPoi (cell) robust GLM (cell) NB GLMM
1.00 0.97 0.93 0.90 0.92 0.89
0.97 1.00 0.90 0.85 0.89 0.85
0.93 0.90 1.00 0.84 0.86 0.85
0.90 0.85 0.84 1.00 0.95 0.96
0.92 0.89 0.86 0.95 1.00 0.93
0.89 0.85 0.85 0.96 0.93 1.00
glmGamPoi (Pb) edgeR (Pb) limma (Pb) glmGamPoi (cell) robust GLM (cell) NB GLMM

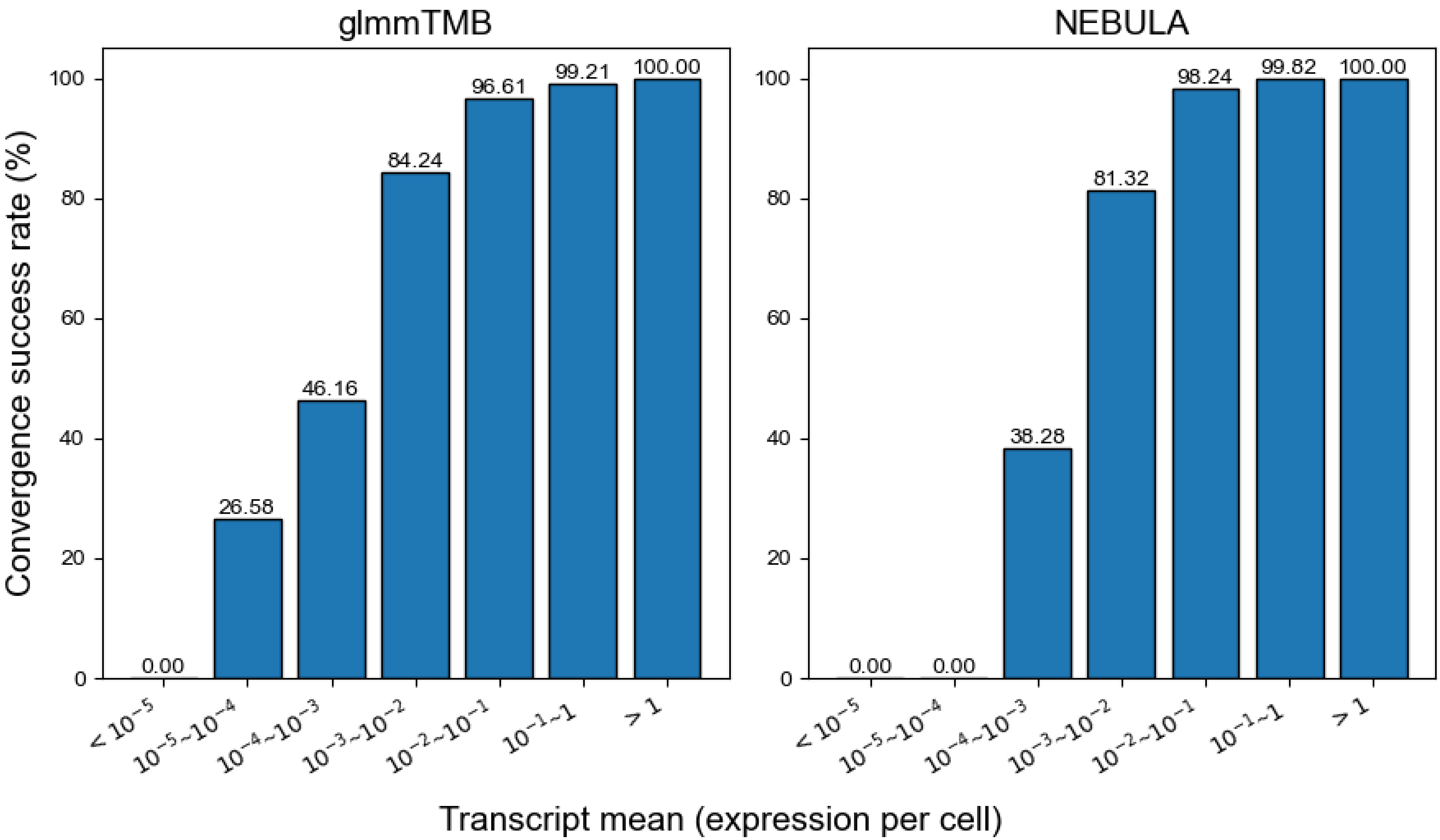
Supplementary Figures

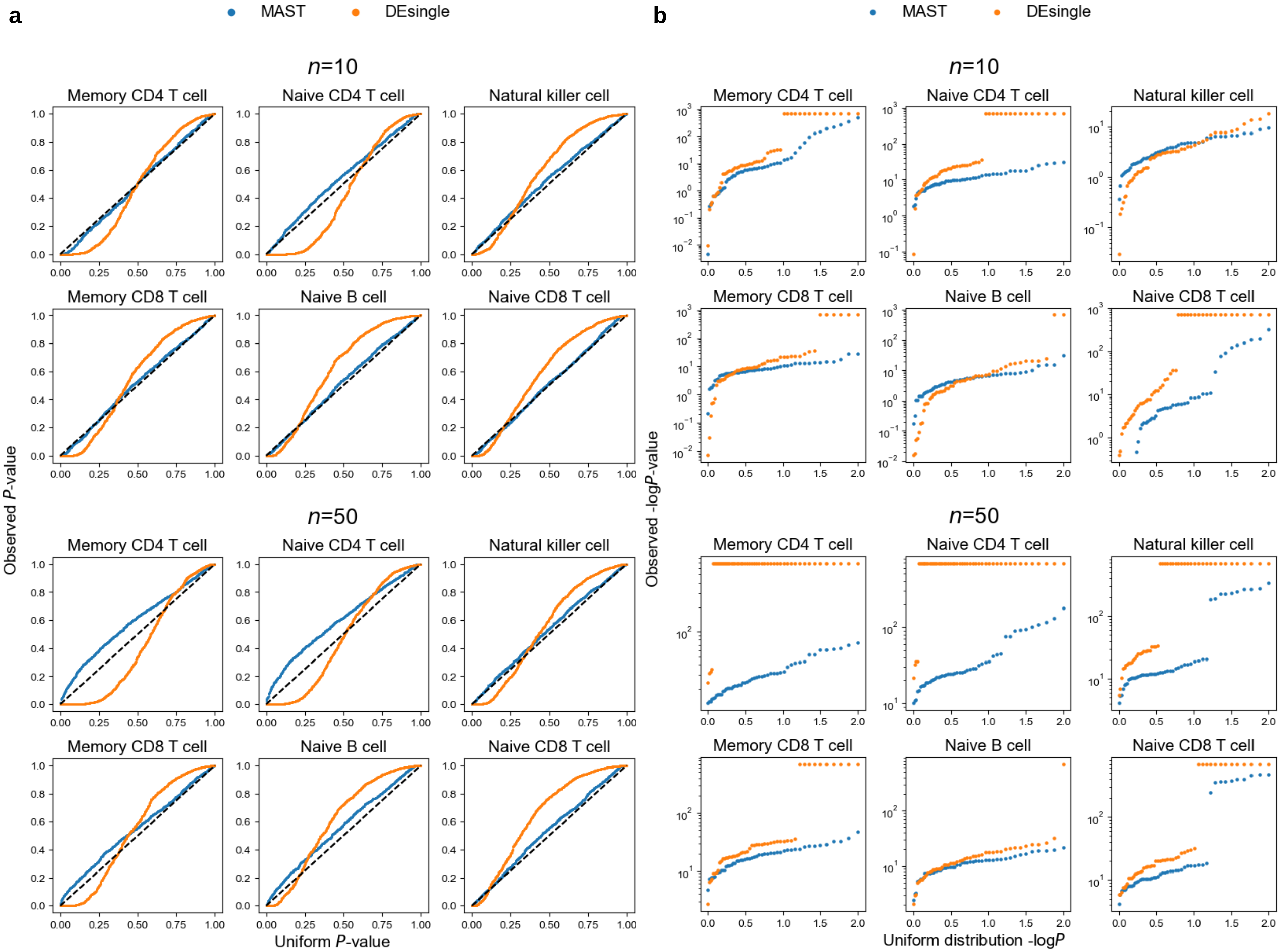

a
glmGamPoi (Pb)
edgeR (Pb)
limma (Pb)
glmGamPoi (cell)
robust GLM (cell)
NB GLMM
n=10
Memory CD4 T cell
Naive CD4 T cell
Natural killer cell
Memory CD8 T cell
Naive B cell
Naive CD8 T cell
n=50
Memory CD4 T cell
Naive CD4 T cell
Natural killer cell
Memory CD8 T cell
Naive B cell
Naive CD8 T cell
Observed P-value
Uniform P-value
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.25
0.50
0.75
1.00
b
glmGamPoi (Pb)
edgeR (Pb)
limma (Pb)
glmGamPoi (cell)
robust GLM (cell)
NB GLMM
n=10
Memory CD4 T cell
Naive CD4 T cell
Natural killer cell
Memory CD8 T cell
Naive B cell
Naive CD8 T cell
n=50
Memory CD4 T cell
Naive CD4 T cell
Natural killer cell
Memory CD8 T cell
Naive B cell
Naive CD8 T cell
Observed -logP-value
Uniform distribution -logP
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
0.0
0.5
1.0
1.5
2.0

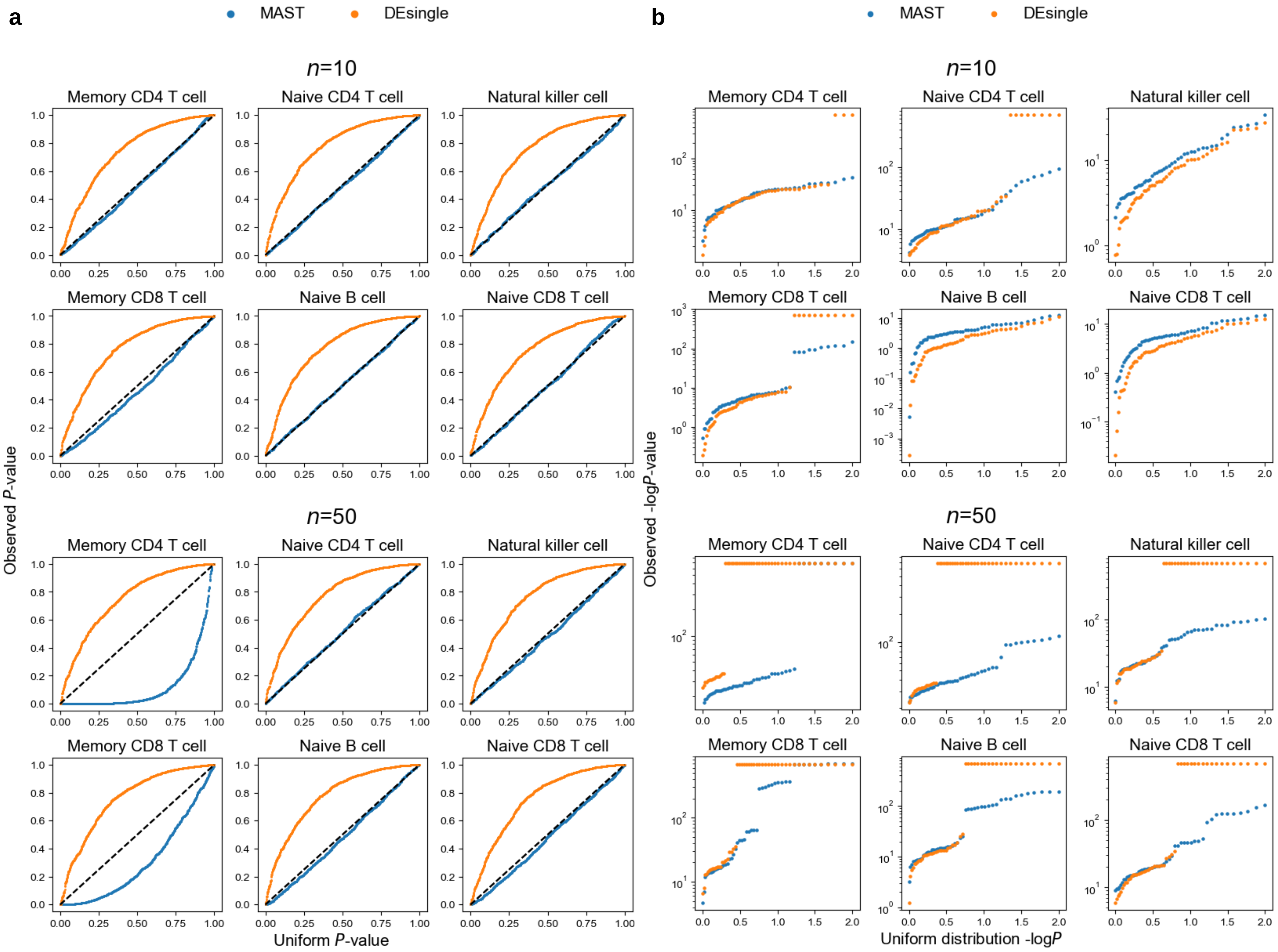

a
glmGamPoi (Pb)
limma (Pb)
robust GLM (cell)
edgeR (Pb)
glmGamPoi (cell)
NB GLMM
n=10
Cardiomyocyte
Mural cell
Fibroblast
Endothelial cell
Myeloids
Native cell
n=50
Cardiomyocyte
Mural cell
Fibroblast
Endothelial cell
Myeloids
Native cell
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.25
0.50
0.75
1.00
Observed P-value
Uniform P-value
b
glmGamPoi (Pb)
limma (Pb)
robust GLM (cell)
edgeR (Pb)
glmGamPoi (cell)
NB GLMM
n=10
Cardiomyocyte
Mural cell
Fibroblast
Endothelial cell
Myeloids
Native cell
n=50
Cardiomyocyte
Mural cell
Fibroblast
Endothelial cell
Myeloids
Native cell
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
0.0
0.5
1.0
1.5
2.0
Observed -logP-value
Uniform distribution -logP

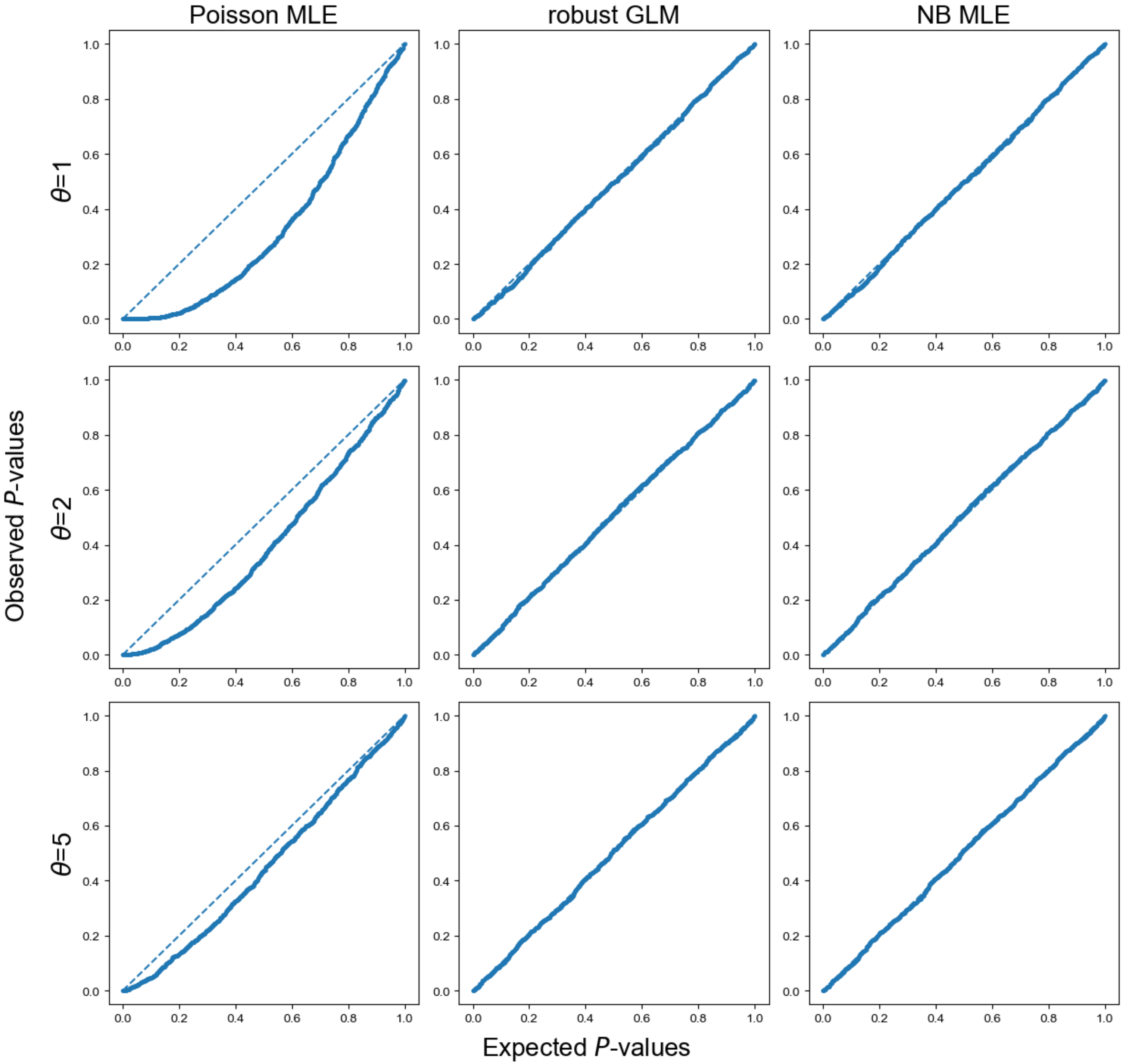

a
glmGamPoi (Pb)
edgeR (Pb)
limma (Pb)
glmGamPoi (cell)
robust GLM (cell)
NB GLMM
n=10
Memory CD4 T cell
Naive CD4 T cell
Natural killer cell
Memory CD8 T cell
Naive B cell
Naive CD8 T cell
n=50
Memory CD4 T cell
Naive CD4 T cell
Natural killer cell
Memory CD8 T cell
Naive B cell
Naive CD8 T cell
Observed P-value
Uniform P-value
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.25
0.50
0.75
1.00
b
glmGamPoi (Pb)
edgeR (Pb)
limma (Pb)
glmGamPoi (cell)
robust GLM (cell)
NB GLMM
n=10
Memory CD4 T cell
Naive CD4 T cell
Natural killer cell
Memory CD8 T cell
Naive B cell
Naive CD8 T cell
n=50
Memory CD4 T cell
Naive CD4 T cell
Natural killer cell
Memory CD8 T cell
Naive B cell
Naive CD8 T cell
Observed -logP-value
Uniform distribution -logP
10^{-1}
10^{0}
10^{1}
10^{2}
0.0
0.5
1.0
1.5
2.0

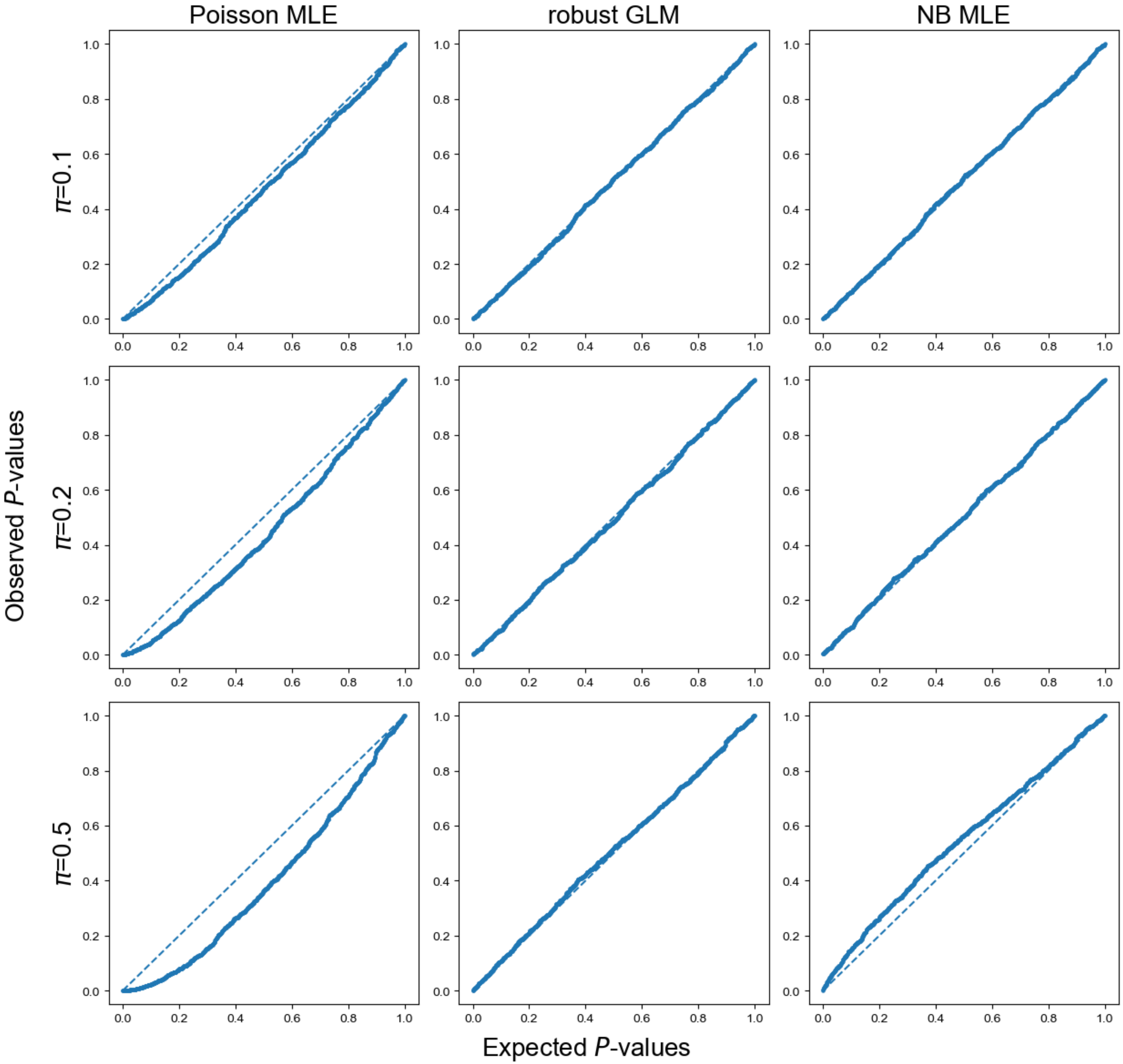

a
● glmGamPoi (Pb)
★ edgeR (Pb)
+ limma (Pb)
■ glmGamPoi (cell)
× robust GLM (cell)
Y NB GLMM
n=10
Cardiomyocyte
Mural cell
Fibroblast
Endothelial cell
Myeloids
Native cell
n=50
Cardiomyocyte
Mural cell
Fibroblast
Endothelial cell
Myeloids
Native cell
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.25
0.50
0.75
1.00
Observed P-value
Uniform P-value
b
● glmGamPoi (Pb)
★ edgeR (Pb)
+ limma (Pb)
■ glmGamPoi (cell)
× robust GLM (cell)
Y NB GLMM
n=10
Cardiomyocyte
Mural cell
Fibroblast
Endothelial cell
Myeloids
Native cell
n=50
Cardiomyocyte
Mural cell
Fibroblast
Endothelial cell
Myeloids
Native cell
10^1
10^2
0.0
0.5
1.0
1.5
2.0
Observed -logP-value
Uniform distribution -logP

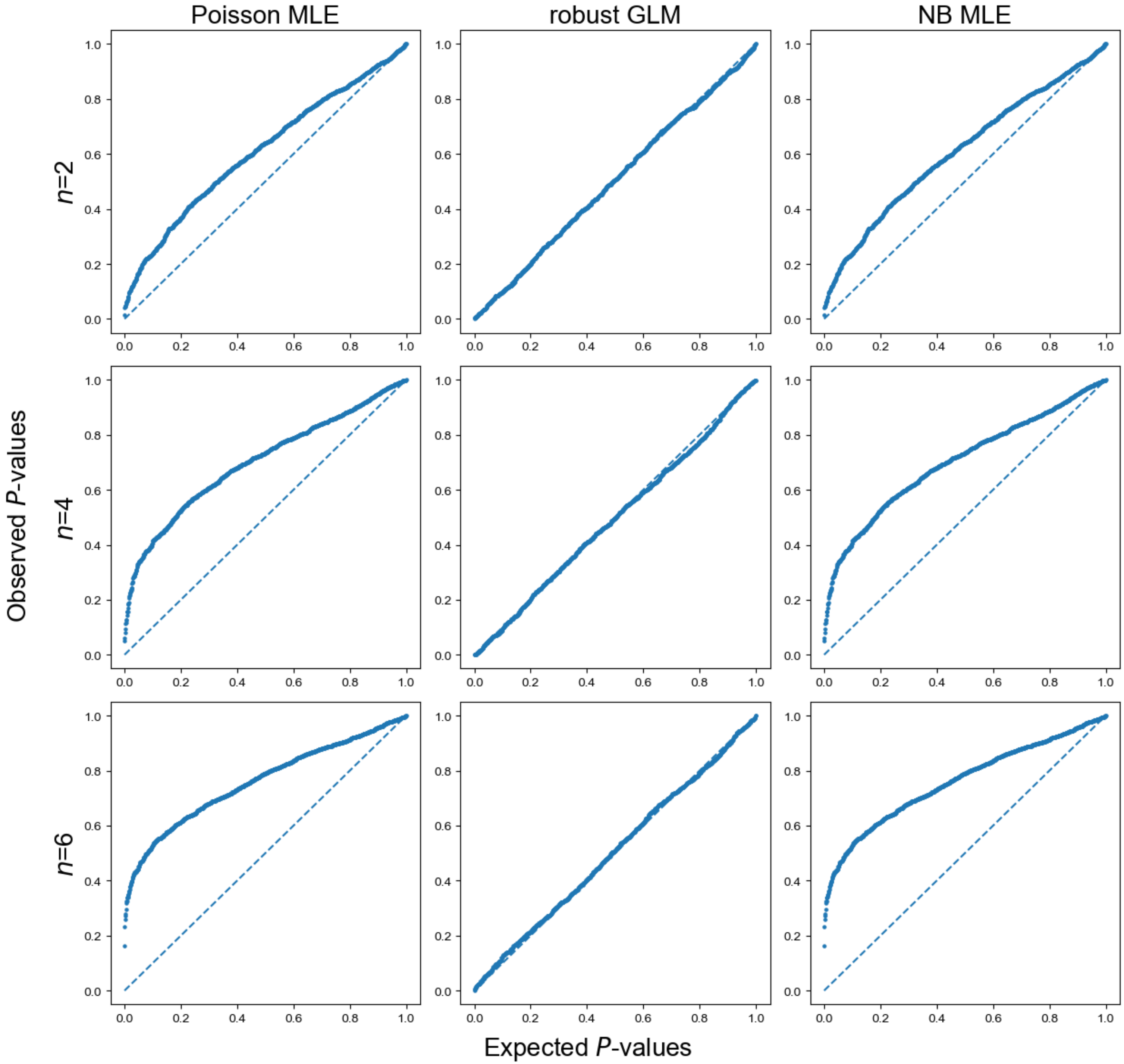

glmmTMB
NEBULA
Convergence success rate (%)
0
20
40
60
80
100
0.00
26.58
46.16
84.24
96.61
99.21
100.00
0.00
0.00
38.28
81.32
98.24
99.82
100.00
$< 10^{-5}$
10^{-5}~10^{-4}
10^{-4}~10^{-3}
10^{-3}~10^{-2}
10^{-2}~10^{-1}
10^{-1}~1
> 1
Transcript mean (expression per cell)


a
● MAST
● DEsingle
n=10
Memory CD4 T cell
Naive CD4 T cell
Natural killer cell
Memory CD8 T cell
Naive B cell
Naive CD8 T cell
n=50
Memory CD4 T cell
Naive CD4 T cell
Natural killer cell
Memory CD8 T cell
Naive B cell
Naive CD8 T cell
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.25
0.50
0.75
1.00
Observed P-value
Uniform P-value
b
● MAST
● DEsingle
n=10
Memory CD4 T cell
Naive CD4 T cell
Natural killer cell
Memory CD8 T cell
Naive B cell
Naive CD8 T cell
n=50
Memory CD4 T cell
Naive CD4 T cell
Natural killer cell
Memory CD8 T cell
Naive B cell
Naive CD8 T cell
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
0.0
0.5
1.0
1.5
2.0
Observed -logP-value
Uniform distribution -logP


a
MAST
DEsingle
n=10
Memory CD4 T cell
Naive CD4 T cell
Natural killer cell
Memory CD8 T cell
Naive B cell
Naive CD8 T cell
n=50
Memory CD4 T cell
Naive CD4 T cell
Natural killer cell
Memory CD8 T cell
Naive B cell
Naive CD8 T cell
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.25
0.50
0.75
1.00
Observed P-value
Uniform P-value
b
MAST
DEsingle
n=10
Memory CD4 T cell
Naive CD4 T cell
Natural killer cell
Memory CD8 T cell
Naive B cell
Naive CD8 T cell
n=50
Memory CD4 T cell
Naive CD4 T cell
Natural killer cell
Memory CD8 T cell
Naive B cell
Naive CD8 T cell
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
0.0
0.5
1.0
1.5
2.0
Observed -logP-value
Uniform distribution -logP


Poisson MLE
robust GLM
NB MLE
θ=1
θ=2
θ=5
Observed P-values
Expected P-values
0.0
0.2
0.4
0.6
0.8
1.0


Poisson MLE
robust GLM
NB MLE
π=0.1
π=0.2
π=0.5
Observed P-values
Expected P-values
0.0
0.2
0.4
0.6
0.8
1.0


Poisson MLE
robust GLM
NB MLE
n=2
n=4
n=6
Observed P-values
Expected P-values
0.0
0.2
0.4
0.6
0.8
1.0

Supplementary Note v2bh

Hanbin Lee

November 6, 2023

1 Setup

We assume that the expression counts obey the following equation.

$$g\left(\mathbb{E}\left[Y_{ijk} \mid X_{ij}, \alpha_{ik}\right]\right) = \beta_{0k} + X_{ij}\beta_{1k} + \alpha_{ik} \tag{1}$$

i is the index of subjects, j is the index of cells, and k is the index of genes. Y_{ijk} is the expression count, X_{ij} is the cell status (0 or 1 for the binary state to compare), and α_{ik} is the subject-specific effect on expression. g is the link function. Alternatively, the equation can be written as

$$Y_{ijk} = g^{-1}\left(\beta_{0k} + X_{ij}\beta_{1k} + \alpha_{ik}\right) + \epsilon_{ijk} \tag{2}$$

where $\epsilon_{ijk} = Y_{ijk} - \mathbb{E}\left[Y_{ijk} \mid X_{ij}, \alpha_{ik}\right]$. Note that $\mathbb{E}\left[\epsilon_{ijk} \mid X_{ij}\right] = 0$.

The link function g is defined based on the assumed model. Linear models preceded by log-like normalization (e.g., limma) follow the above equation with $g(x) = x$, the identity function. Poisson regression, negative binomial regression, and their associated mixed models use $g(x) = \log x$, the log function. For the sake of generality, the following derivations do not assume a particular probability distribution.

When $g(x) = \log x$, β_{1k} is the log fold-change (logFC) which is the parameter of interest. Even for the $g(x) = x$ case, after suitable log transformation of Y_{ijk}, β_{1k} is an approximation of the logFC. In the following sections, the proofs assume $g(x) = \log x$ because the derivation is more difficult than $g(x) = x$. The proofs for the $g(x) = x$ case can be obtained from the $g(x) = \log x$ case after removing all exp and log functions.

2 Autocorrelation and pseudoreplication

When the expressions in the cells within a subject are correlated (not independent), we call it *autocorrelation.* In equation (1), α_{ik} is the term responsible for the autocorrelation. As α_{ik} varies across subjects, and subjects are randomly selected from the population, α_{ik} is a random variable with non-zero variance. If it were constant, the intercept β_{0k} absorbs the term leaving no autocorrelation. Therefore, autocorrelation can only exist when there are two or more subjects since all cells will have the same α_{ik} when there is only one subject. We assume that autocorrelation exists in the data (otherwise, everything simplifies, and applying the cell-wise method to any data should work).

In literature, autocorrelation is attributed as the cause of *pseudoreplication bias* [3, 2]. There are several observations related to the bias. First, the data should have of multiple subjects so that autocorrelation is present. Second, when a cell-wise method is applied to a multi-subject data, the false discovery rate (FDR) increases substantially. Third, aggregating the cells into a pseudobulk ameliorates the inflated FDR. Another alternative is to use a mixed model. It follows that autocorrelation causes the bias, and the two methods resolve it by taking autocorrelation into account. Following the same logic, the failure of cell-wise methods comes from ignoring autocorrelation.

Autocorrelation explains the name of the bias. *Replicate* refers to the independent units of observation from a study [1, 3]. As cells are mutually correlated, they are not replicates. Instead, they are called *pseudoreplicates* for not being independent. Since cell-wise methods treat cells (which are pseudoreplicates) as independent observations (i.e. replicates), it makes sense to call the high FDR of cell-wise methods pseudoreplication bias because it stems from mistaking pseudoreplicates as independent replicates.

The argument gives the impression that pseudoreplication bias happens whenever cell-wise method is used in the presence of autocorrelation. However, we have shown that cell-wise methods are well-calibrated in the second scenario discussed in the main text. To close the gap between the previous understanding of pseudoreplication bias and our observations, we suggest a more narrow and precise definition of pseudoreplication bias based on theoretical speculation.

3 Pseudoreplication bias revisited

We show below that the pseudoreplication bias occurs only if the three conditions are all met: (1) multi-subject design (with autocorrelation), (2) constant cell states within a subject (namely scenario 1), and (3) cell-wise method used. We recast the bias in the language of estimator variance to analytically prove our argument.

An estimator $\widehat{\beta}_{1k}$ guesses the parameter β_{1k} with a finite amount of data. Since the estimator is a function of the data which is a random variable, it is also random. It means that the estimator varies to a certain degree for each experiment. An estimator's variance measures the variability of the estimator under repeated (but identical) experiments. Small variance means less variability, i.e. the value of the estimator is likely to concentrate in a smaller range under repeated experiments. However, we usually have a single realization of potentially many experiments so the true variance is unknown. The estimated variance is then used to compute the P-value to assess statistical significance. In short, there is the true variance of an estimator under repeated experiments, $\operatorname{Var}\left(\widehat{\beta}_{1k}\right)$, and the estimated variance $\widehat{\operatorname{Var}}\left(\widehat{\beta}_{1k}\right)$ that is used in practice because the true variance is only theoretically known.

A method to detect DEG is equipped with an estimator $\widehat{\beta}_{1k}$ and its corresponding variance estimate $\widehat{\operatorname{Var}}\left(\widehat{\beta}_{1k}\right)$. If the estimate $\widehat{\operatorname{Var}}\left(\widehat{\beta}_{1k}\right)$ is a proper guess for $\operatorname{Var}\left(\widehat{\beta}_{1k}\right)$, the test is valid and produces calibrated P-values. Conversely, if the method underestimates the true variance, it produces deflated P-values, leading to an inflation of false discoveries. For example, z-test computes

$$Z = \frac{\widehat{\beta}_{1k}}{\widehat{\operatorname{Var}}\left(\widehat{\beta}_{1k}\right)} \tag{3}$$

and tests if $|Z|$ exceeds a pre-specified threshold to declare significance. Thus, an underestimation of $\widehat{\operatorname{Var}}\left(\widehat{\beta}_{1k}\right)$ increases $|Z|$ leading to more false discoveries even if $\widehat{\beta}_{1k}$ is unbiased. Other tests like χ^2-test, Wald test, and likelihood ratio test (LRT) also utilize the similar variance estimate. Thus, the significance is exaggerated whenever the variance is underestimated, regardless of the test method used.

For any cell-wise method, the estimated variance of its estimator is always

$$\widehat{\operatorname{Var}}\left(\widehat{\beta}_{1k}\right) = \mathcal{O}\left(\frac{1}{N}\right) \tag{4}$$

where N is the number of cells, which is a consequence of assuming that N cells are independent observations (replicates). Unfortunately, the true variance has different scales depending on the study design. Thus, the scale of the true variance and the estimated variance may not coincide.

Theorem 1.

$$\operatorname{Var}\left(\widehat{\beta}_{1k}\right) = \begin{cases} \mathcal{O}\left(\frac{1}{n}\right) & : X_{ij} \textit{ is constant within } i\textit{, Scenario 1} \\ \mathcal{O}\left(\frac{1}{N}\right) & : X_{ij} \textit{ varies within } i\textit{, Scenario 2} \end{cases} \tag{5}$$

for any unbiased estimator $\widehat{\beta}_{1k}$ for the true logFC (β_{1k}). n and N are the number of subjects and cells, respectively.

The theorem shows that the variance estimate of any cell-wise method (4) is a severe underestimate of the true variance under scenario 1, which should increase the chance of false discovery to a large extent. This matches our previous understanding of pseudoreplication bias but with more restrictive conditions: FDR increases when cell-wise method is used in a multi-subject study but only in the first scenario where the underestimation of variance happens.

Corollary 1. *Pseudoreplication bias is the inflation of FDR produced by cell-wise methods in scenario 1. The inflated FDR comes from the underestimation of the estimator's variance. The following equation gives the degree of underestimation.*

$$\widehat{\operatorname{Var}}\left(\widehat{\beta}_{1k}\right) = \mathcal{O}\left(\frac{1}{N}\right) \ll \mathcal{O}\left(\frac{1}{n}\right) = \operatorname{Var}\left(\widehat{\beta}_{1k}\right) \tag{6}$$

since $n \ll N$.

Theorem 1 also explains why pseudobulk methods are an adequate solution to the inflated FDR in scenario 1. Pseudobulk methods treat subjects as independent replicates. Therefore, the estimated variance of the estimator has the following scale

$$\widehat{\operatorname{Var}}\left(\widehat{\beta}_{1k}\right) = \mathcal{O}\left(\frac{1}{n}\right) \tag{7}$$

which matches the scale of the true variance of $\widehat{\beta}_{1k}$.

Finally, the calibrated performance of cell-wise methods in the second scenario can be understood. The scale of cell-wise methods in equation (4) equals that of the true variance in the second scenario $\mathcal{O}(1/N)$. This implies that cells are true replicates in the second scenario. Therefore, autocorrelation alone is not enough to determine the true independent unit of a study.

Corollary 2. *The scale of the estimator variance of cell-wise methods equals to the scale of the true variance under scenario 2.*

$$\widehat{\mathrm{Var}}\left(\widehat{\beta}_{1k}\right) = \mathcal{O}\left(\frac{1}{N}\right) = \mathrm{Var}\left(\widehat{\beta}_{1k}\right) \tag{8}$$

Note that the equality sign '=' here means they coincide in terms of scale, not being numerically the same.

Now we move on to the proofs of **Theorem 1**.

Proof. The proof of the claim (**Theorem 1**) comes in two steps. First, we analyze the behavior of $\widehat{\beta}_{1k}$ conditioned on the sampled subjects. This fixes $\boldsymbol{\alpha}_k$, the collection of all α_{ik}, removing the autocorrelation between the cells. Next, we let $\boldsymbol{\alpha}_k$ vary and observe the full behavior. The law of iterated variance gives

$$\mathrm{Var}\left(\widehat{\beta}_{N,1k}\right) = \mathrm{Var}\left(\mathbb{E}\left[\widehat{\beta}_{N,1k} \mid \mathbf{X}, \boldsymbol{\alpha}_k\right]\right) + \mathbb{E}\left[\mathrm{Var}\left(\widehat{\beta}_{N,1k} \mid \mathbf{X}, \boldsymbol{\alpha}_k\right)\right] \tag{9}$$

where $\mathbf{X}$ is the collection of all X_{ij} (cell status of cell j of individual i). The subscript N means the estimator is computed from N cells. Conditioning on $\mathbf{X}$ and $\boldsymbol{\alpha}_k$ gives the effect of fixing the sampled subjects. As explained earlier, we analyze $\mathbb{E}\left[\widehat{\beta}_{N,1k} \mid \mathbf{X}, \boldsymbol{\alpha}_k\right]$ and $\mathrm{Var}\left(\widehat{\beta}_{N,1k} \mid \mathbf{X}, \boldsymbol{\alpha}_k\right)$ first which describe the behavior of the estimator with $\mathbf{X}$ and $\boldsymbol{\alpha}_k$ fixed.

Proposition 1. *The behavior of the second term,* $\mathrm{Var}\left(\widehat{\beta}_{N,1k} \mid \mathbf{X}, \boldsymbol{\alpha}_k\right)$*, is design-independent, and*

$$\mathrm{Var}\left(\widehat{\beta}_{N,1k} \mid \mathbf{X}, \boldsymbol{\alpha}_k\right) = \mathcal{O}\left(\frac{1}{N}\right) \tag{10}$$

Proof. As $\mathbf{X}$ and $\boldsymbol{\alpha}_k$ are fixed, the cells are independent replications. This is because the only remaining variation of $\widehat{\beta}_{N,1k}$ after fixing $\mathbf{X}$ and $\boldsymbol{\alpha}_k$ comes from the independent sampling of cells within each subject. Then the central limit theorem of independent observation of cells gives the desired result as $N \to \infty$.

⊠

Subsequently,

$$\mathbb{E}\left[\operatorname{Var}\left(\widehat{\beta}_{N,1k} \mid \mathbf{X}, \boldsymbol{\alpha}_k\right)\right] = \mathcal{O}\left(\frac{1}{N}\right) \tag{11}$$

since the expectation of an $\mathcal{O}(1/N)$ variable is $\mathcal{O}(1/N)$.

The interesting part that exhibits design-dependent behavior is the first term, $\operatorname{Var}\left(\mathbb{E}\left[\widehat{\beta}_{N,1k} \mid \mathbf{X}, \boldsymbol{\alpha}_k\right]\right)$. The conditional expectation, $\mathbb{E}\left[\widehat{\beta}_{N,1k} \mid \mathbf{X}, \boldsymbol{\alpha}_k\right]$ is the expected value of $\widehat{\beta}_{N,1k}$ provided the subjects are fixed. It is difficult to directly expand this quantity. However, it is doable if we consider the conditional mean expression of a specific cell. Under the condition with $\mathbf{X}$ and $\boldsymbol{\alpha}_k$ fixed, assume that we randomly select one cell (cell j' of subject i'). If we consider the expected expression of gene k of this cell:

$$\begin{aligned}
&\mathbb{E}\left[Y_{i'j'k} \mid X_{i'j'}, \mathbf{X}, \boldsymbol{\alpha}_k\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_{i'j'k} \mid X_{i'j'}, \alpha_{i'k}, \mathbf{X}, \boldsymbol{\alpha}_k\right] \mid X_{i'j'}, \mathbf{X}, \boldsymbol{\alpha}_k\right] \quad \text{(law of total expectation)} \\
&= \mathbb{E}\left[\exp\left(\beta_{0k} + X_{i'j'}\beta_{1k} + \alpha_{i'k}\right) \mid X_{i'j'}, \mathbf{X}, \boldsymbol{\alpha}_k\right] \quad \text{(by the definition of } Y_{i'j'k} \text{ in eq. (1))} \\
&= \exp\left(\beta_{0k} + X_{i'j'}\beta_{1k}\right) \cdot \mathbb{E}\left[\exp\left(\alpha_{i'k}\right) \mid X_{i'j'}, \mathbf{X}, \boldsymbol{\alpha}_k\right] \quad \text{(constants pulled out)} \\
&= \exp\left(\beta_{0k} + X_{i'j'}\beta_{1k} + \log \mathbb{E}\left[\exp\left(\alpha_{i'k}\right) \mid X_{i'j'}, \mathbf{X}, \boldsymbol{\alpha}_k\right]\right)
\end{aligned} \tag{12}$$

We further analyze $\mathbb{E}\left[\exp\left(\alpha_{i'k}\right) \mid X_{i'j'}, \mathbf{X}, \boldsymbol{\alpha}_k\right]$. This is the design-dependent term of interest. Let $S_{i'j'} = 1, \ldots, n$ be the subject membership of cell j'. (That is, $S_{i'j'} = i'$. However the proof will look easier to read with this.) $\mathbb{I}(\,\cdot\,)$ is the indicator function.

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\alpha_{i'k}\right) \mid X_{i'j'}, \mathbf{X}, \boldsymbol{\alpha}_k\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\alpha_{i'k}\right) \mid S_{i'j'}, X_{i'j'}, \mathbf{X}, \boldsymbol{\alpha}_k\right] \mid X_{i'j'}, \mathbf{X}, \boldsymbol{\alpha}_k\right] \quad \text{(law of total expectation)} \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\alpha_{i'k}\right) \mid S_{i'j'}, \mathbf{X}, \boldsymbol{\alpha}_k\right] \mid X_{i'j'}, \mathbf{X}, \boldsymbol{\alpha}_k\right] \quad (X_{i'j'} \perp \exp(\alpha_{i'k}) \mid S_{i'j'}) \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \exp(\alpha_{ik}) \mathbb{I}(S_{i'j'} = i) \,\middle|\, X_{i'j'}, \mathbf{X}, \boldsymbol{\alpha}_k\right] \quad (\alpha_{ik} \text{ is constant given } S_{i'j'}) \\
&= \sum_{i=1}^{n} \exp(\alpha_{ik}) \cdot \mathbb{E}\left[\mathbb{I}(S_{i'j'} = i) \mid X_{i'j'}, \mathbf{X}, \boldsymbol{\alpha}_k\right] \\
&= \sum_{i=1}^{n} \exp(\alpha_{ik}) \cdot \mathbb{P}(S_{i'j'} = i \mid X_{i'j'}, \mathbf{X}, \boldsymbol{\alpha}_k) \quad \text{(expectation of an indicator is probability)}
\end{aligned} \tag{13}$$

In scenario 1, $X_{i'j'} = X_{i'}$ because the cells from subject i' all have the same value. Given $X_{i'} = 1$, $\alpha_{i'k}$ is sampled from $\boldsymbol{\alpha}_k$ such that $X_i = 1$, and vice versa. Therefore,

$$\mathbb{P}(S_{i'j'} = i \mid X_{i'} = x, \mathbf{X}, \boldsymbol{\alpha}_k) = \begin{cases} N_i/(\sum_{h:X_h=x} N_h) & : X_i = x \\ 0 & : X_i \neq x \end{cases} \tag{14}$$

which is the probability among subjects $X_i = x$ or zero, otherwise. N_i is the number of cells in subject i, such that $N = N_1 + ... + N_n$.

In scenario 2, the fact that $X_{i'j'}$ varies independently within a subject implies $S_{i'j'} \perp X_{i'j'} \mid \mathbf{X}, \boldsymbol{\alpha}$. Hence,

$$\mathbb{P}(S_{i'j'} = i \mid X_{i'j'} = x, \mathbf{X}, \boldsymbol{\alpha}_k) = \mathbb{P}(S_{i'j'} = i \mid \mathbf{X}, \boldsymbol{\alpha}_k) = \frac{N_i}{\sum_{h=1}^n N_h} \tag{15}$$

which is the probability of cell j' coming from subject i. The difference is that equation (14) depends on $X_{i'j'}$ while (15) doesn't. For later use, let $p_{i|x} = \mathbb{P}(S_{i'j'} = i \mid X_{i'} = x, \mathbf{X}, \boldsymbol{\alpha}_k)$ and $p_i = \mathbb{P}(S_{i'j'} = i \mid \mathbf{X}, \boldsymbol{\alpha}_k)$.

Let $M_0 = \sum_{i=1}^n \exp(\alpha_{ik}) p_{i|0}$, $M_1 = \sum_{i=1}^n \exp(\alpha_{ik}) p_{i|1}$, and $M = \sum_{i=1}^n \exp(\alpha_{ik}) p_i$. Substituting equation (14) and (15) to log of equation (13) gives

$$\begin{aligned} &\log \mathbb{E}\left[\exp\left(\alpha_{i'k}\right) \mid X_{i'j'}, \mathbf{X}, \boldsymbol{\alpha}_k\right] \\ &= \begin{cases} \left(\log M_1 - \log M_0\right) X_{i'j'} + \log M_0 & : \text{Scenario 1} \\ \log M & : \text{Scenario 2} \end{cases} \end{aligned} \tag{16}$$

Substituting equation (16) to equation (12) gives

$$\begin{aligned} &\mathbb{E}\left[Y_{i'j'k} \mid X_{i'j'}, \alpha_{i'k}, \mathbf{X}, \boldsymbol{\alpha}_k\right] \\ &= \exp\left(\beta_{0k} + X_{i'j'}\beta_{1k} + \log \mathbb{E}\left[\exp\left(\alpha_{i'k}\right) \mid X_{i'j'}, \mathbf{X}, \boldsymbol{\alpha}_k\right]\right) \\ &= \begin{cases} \exp\left(\log M_0 + \beta_{0k} + X_{i'j'}[\beta_{1k} + \log M_1 - \log M_0]\right) & : \text{Scenario 1} \\ \exp\left(\log M + \beta_{0k} + X_{i'j'}\beta_{1k}\right) : \text{Scenario 2} \end{cases} \end{aligned} \tag{17}$$

Thus, the expectation of the estimator $\widehat{\beta}_{1k}$ conditional on $\mathbf{X}$ and $\boldsymbol{\alpha}$ is

$$\mathbb{E}\left[\widehat{\beta}_{N,1k} \mid \mathbf{X}, \boldsymbol{\alpha}_k\right] = \begin{cases} \beta_{1k} + \log M_1 - \log M_0 & : \text{Scenario 1} \\ \beta_{1k} & : \text{Scenario 2} \end{cases} \tag{18}$$

because these are the coefficients of X_{ij} in equation (17).

Subsequently, the variance is

$$\text{Var}\left(\mathbb{E}\left[\widehat{\beta}_{N,1k} \mid \mathbf{X}, \boldsymbol{\alpha}_k\right]\right) = \begin{cases} \text{Var}\left(\log M_1 - \log M_0\right) & : \text{Scenario 1} \\ 0 & : \text{Scenario 2} \end{cases} \tag{19}$$

The expansion of $\text{Var}\left(\log M_1 - \log M_0\right)$ gives

$$\begin{aligned} &\text{Var}\left(\log M_1 - \log M_0\right) \\ &= \text{Var}\left(\log M_1\right) + \text{Var}\left(\log M_0\right) \quad \text{(sampling of subjects is independent)} \\ &= \text{Var}\left(\log\left[\sum_{i:X_i=1} \exp(\alpha_{ik}) p_{i|1}\right]\right) + \text{Var}\left(\log\left[\sum_{i:X_i=0} \exp(\alpha_{ik}) p_{i|0}\right]\right) \end{aligned} \tag{20}$$

Since $p_{i|x} \sim 1/n$ for all subjects i (it means that the number of cells from the subjects are in a similar order), the scale of M_x $(x = 0, 1)$ is

$$\mathcal{O}(M_x) = \mathcal{O}\left(\frac{1}{n} \sum_{i:\text{all subjects}} \exp(\alpha_{ik})\right) = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \tag{21}$$

by the central limit theorem applied to independent subjects. Therefore, the variance of the $\log M_x$ is $\mathcal{O}(1/n)$.

Now we arrive at our main conclusion.

$$\begin{aligned}
\text{Var}\left(\widehat{\beta}_{N,1k}\right) &= \text{Var}\left(\mathbb{E}\left[\widehat{\beta}_{N,1k} \mid \mathbf{X}, \boldsymbol{\alpha}_k\right]\right) + \mathbb{E}\left[\text{Var}\left(\widehat{\beta}_{N,1k} \mid \mathbf{X}, \boldsymbol{\alpha}_k\right)\right] \\
&= \begin{cases} \mathcal{O}\left(\frac{1}{n}\right) + \mathcal{O}\left(\frac{1}{N}\right) & : \text{Scenario 1} \\ 0 + \mathcal{O}\left(\frac{1}{N}\right) & : \text{Scenario 2} \end{cases} \\
&= \begin{cases} \mathcal{O}\left(\frac{1}{n}\right) & : \text{Scenario 1} \\ \mathcal{O}\left(\frac{1}{N}\right) & : \text{Scenario 2} \end{cases}
\end{aligned} \tag{22}$$

⊠

Theorem 1 clarifies the name of pseudoreplication bias. Virtually all tests, if not most, rely on the central limit theorem to test significance. A normal distribution approximation is applied at a certain step in the procedure to obtain P-values. Loosely speaking, the goodness of the approximation depends on the number of effective sample size, and this number is used to assess significance. The number also measures the amount of data available for statistical inference. In multi-subject scRNA-seq analysis, this number is ambiguous as there are at least two levels of units: the subjects and the cells. **Theorem 1** resolves the ambiguity by explicating that the number is the number of subjects (n) and the number of cells (N) in each design, respectively. Hence, cells are not the true replicates that drive the normal approximation of the test in the first scenario, so we call them pseudoreplicates. Nevertheless, cell-wise tests always take N as the effective sample size, mistaking the true replicates as the cells. Therefore, it's called the pseudoreplication bias for using pseudoreplicates (cells) instead of true replicates (subjects) for the test.

4 Robust methods under dropouts

Let $Z_{ijk} = 0, 1$ is the dropout indicator variable that is 1 when dropout occurs and 0 when it doesn't. Then, a zero-inflated count W_{ijk} is

$$W_{ijk} = Z_{ijk} \cdot 0 + (1 - Z_{ijk}) \cdot Y_{ijk} \tag{23}$$

where Y_{ijk} follows the distribution that obeys the conditional mean equation (1).

Assuming that the dropout event occurs independent of Y_{ijk} conditional on α_{ik} and X_{ij},

$$\begin{aligned}\mathbb{E}\left[W_{ijk} \mid X_{ij}, \alpha_{ik}\right] &= \mathbb{E}\left[(1 - Z_{ijk})Y_{ijk} \mid X_{ij}, \alpha_{ik}\right] \\ &= \mathbb{E}\left[(1 - Z_{ijk}) \mid X_{ijk}, \alpha_{ik}\right] \cdot \mathbb{E}\left[Y_{ijk} \mid X_{ij}, \alpha_{ik}\right]\end{aligned} \tag{24}$$

which implies

$$\begin{aligned}\log \mathbb{E}\left[W_{ijk} \mid X_{ij}, \alpha_{ik}\right] &= \log \mathbb{E}\left[(1 - Z_{ijk}) \mid X_{ij}, \alpha_{ik}\right] + \log \mathbb{E}\left[Y_{ijk} \mid X_{ij}, \alpha_{ik}\right] \\ &= \log \mathbb{E}\left[(1 - Z_{ijk}) \mid X_{ij}, \alpha_{ik}\right] + \beta_{0k} + X_i\beta_{1k} + \alpha_{ik}\end{aligned} \tag{25}$$

Since X_{ij} is discrete and the number of subjects is finite, $\log \mathbb{E}\left[(1 - Z_{ijk}) \mid X_{ij}, \alpha_{ik}\right]$ can be written as a linear function respect to X_{ij} and α_j using dummy variables. Therefore, equation (25) obeys the conditional mean equation (1). Therefore, robust methods only requiring the conditional mean equation remain valid when dropout is present.

References

[1] S. H. Hurlbert. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, 54(2):187–211, June 1984.

[2] J. W. Squair, M. Gautier, C. Kathe, M. A. Anderson, N. D. James, T. H. Hutson, R. Hudelle, T. Qaiser, K. J. E. Matson, Q. Barraud, A. J. Levine, G. L. Manno, M. A. Skinnider, and G. Courtine. Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12(1), Sept. 2021.

[3] K. D. Zimmerman, M. A. Espeland, and C. D. Langefeld. A practical solution to pseudoreplication bias in single-cell studies. *Nature Communications*, 12(1), Feb. 2021.