Benchmarking integration of single-cell differential expression

Hai Nguyen
UNIST https://orcid.org/0000-0002-8385-1075
Bukyung Baik
Ulsan National Institute of Science and Technology
Sora Yoon
University of Pennsylvania
Taesung Park
Dougu Nam (dougnam@unist.ac.kr)
Ulsan National Institute of Science and Technology

Article

Keywords:

Posted Date: June 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1723455/v1

Version of Record: A version of this preprint was published at Nature Communications on March 21st, 2023. See the published version at https://doi.org/10.1038/s41467-023-37126-3.

Abstract

Integration of single-cell RNA sequencing (scRNA-seq) data between different samples has been a major challenge for analyzing cell populations. However, strategies to integrate differential expression (DE) analysis of scRNA-seq data remain underinvestigated. Here, we benchmarked 41 methods for integrative DE analysis of scRNA-seq data. Batch-effects, sparsity of data, and heterogeneity of samples substantially impacted the performance of DE analysis. Several methods that yielded high performances were suggested based on various simulations and real data analyses. In particular, the bulk RNA-seq tool edgeR incorporating the observation weights and the scRNA-seq tool MAST showed overall good performances. Remarkably, analysis for a specific cell type outperformed that of large-scale bulk sample data in prioritizing disease-related genes and pathways.

Introduction

Recent advances in single-cell RNA sequencing (scRNA-seq) techniques have tremendously increased our understanding of cell types and progresses in disease[1,2]. While thousands of cells have been sequenced in individual studies (or samples), integration of scRNA-seq data has been confounded by technical variations between studies, called *batch effects*. In particular, the lack of starting materials in scRNA-seq results in highly sparse and noisy data, posing a great challenge to batch-effect correction (BEC) of scRNA-seq data[3,4]. To address this problem, various BEC algorithms have been developed for accurately discriminating cell types from multiple scRNA-seq datasets[3]. However, the impact of batch effects on gene-based analysis such as differential expression (DE) analysis and the strategies for integrating DE analysis for scRNA-seq data have remained underinvestigated. Accurate DE analysis in each cell type across samples (or patients) is able to suggest important genes and functions to understand the mechanisms of disease.

Tran and colleagues[3] recently benchmarked 14 BEC methods for scRNA-seq data, and recommended several high performance methods. Most of the BEC methods exploit the low dimensionality of data and remove the technical differences between matched cells using deep learning or statistical models. Some methods then return batch-effect-corrected data (BEC data) in the original high dimension for downstream analysis. In particular, they tested the use of BEC data for DE analysis ("bimod" method[5]) under a simple batch condition, where the analysis of BEC data showed a superior performance compared to that of uncorrected data. In contrast, it was suggested that batch alignment could severely distort the high-dimensional observation of genes, making gene-based analyses problematic[6], and DE testing for measured data with technical covariates included in the model was recommended over using BEC data[7].

While BEC methods are used to reduce or eliminate the technical differences between matched cells, they introduce new errors that accompany dimension reduction and estimation of batch differences. Therefore, the possible improvements in DE analysis by using BEC data should be investigated extensively using various DE methods and batch conditions. Contrary to BEC methods, a statistical model

with a batch covariate, denoted as a *covariate model*, uses the uncorrected data in each batch when estimating the model parameters with which DE is tested (see Methods)[8–10]. Another possible approach for integrating scRNA-seq DE analysis is the meta-analysis, where DE analysis is performed for each batch and the resulting statistics or *p*-values are combined for each gene[11,12].

In this study, we compared various DE methods for scRNA-seq data with multiple batches collected from three approaches: (1) DE analysis of BEC data, (2) covariate modeling, and (3) meta-analysis. We considered "paired" DE analysis where each batch contained all the sample conditions to be compared. Both model-based and model-free simulations of scRNA-seq data were used, and the impacts of batch-effects level, sparsity of data, and heterogeneity between samples were analyzed. Furthermore, we counted DE genes that reversed their signs for each DE method to compare the extent of data distortion. Finally, we analyzed real scRNA-seq data for seven patients with lung adenocarcinoma (LUAD). Notably, the analysis of epithelial cells prioritized both known disease genes and prognostic genes significantly better than that of large-scale bulk sample data, demonstrating the high resolution and efficacy of scRNA-seq DE analysis. Pathway enrichment analysis also supported the relevance of the results[31]. From this study, we suggested several methods that performed well in various tests with simulation and real data.

Results

In total, 41 combinations between six BEC methods (ZINB-WaVE[13], MNNCorrect[14], scMerge[15], Seurat v3[16], limma_BEC[10] and ComBat[17]), covariate models, four meta-analysis methods (Fisher, weighted Fisher (wFisher)[12], fixed effects model (FEM)[11] and random effects model (REM)[11]), and eight DE methods (DESeq2[9], edgeR[18], edgeR_DetRate[19], limma[10], limmatrend[20], moderated *t*-test[21], MAST[22] and the Wilcoxon ranksum test) were tested. These 41 methods are denoted as *integrative* (*DE*) *methods*. We note that all the six BEC methods tested here yield BEC data that can be used for DE analysis. Refer to Supplementary Information on how each integrative method was implemented. We focused on comparisons between two cell groups (test *vs.* control groups) and considered two, four, or seven batches. For each integrative method, a threshold of *q*-value < 0.05 was used to select differentially expressed genes (DE genes). For simulated data, F-score and area under precision-recall curve (AUPR) were compared between methods. In particular, we compared $\mathrm{F}_{0.5}$-scores and partial AUPR (denoted as pAUPR) for recall rates < 0.5, both of which weighed precision higher than recall, considering the high noise rate and sparsity of scRNA-seq data (see Methods). For real scRNA-seq data (LUAD), the ranks of known disease genes and prognostic genes, false-positive rates (*p*-value < 0.05), and false discoveries (*q*-value < 0.05) were compared. The real data contained seven batches for which we additionally tested four pseudobulk analysis methods[23]. Throughout this study, unless otherwise stated, sparsely expressed genes (zero rate > 0.95) were filtered for reliable DE analysis. We observed similar trends when a less strict threshold (zero rate > 0.99) was used; however, we focus on the results using the more strict condition, considering that genes that are very rarely expressed in a cell type are less likely to play a crucial role in disease.

Model-based simulation tests

ScRNA-seq data were simulated on the basis of negative binomial (NB) model[24]. Sparse data with a high overall zero rate (> 80%) after the gene filtering were simulated for each batch. We used a rather challenging condition with large batch effects and small group differences to analyze the effects of each integrative strategy (Fig. 1a-b). The sizes of batch effects and group differences were compared using the principal variance component analysis (PVCA)[25]. The $\mathrm{F}_{0.5}$-scores and precision-recall analysis results for the two-batch case are shown in Fig. 1c-d. The experiment was performed for several combinations of "dropout" parameter values and test-control ratios, and the resulting $\mathrm{F}_{0.5}$-scores and pAUPRs were represented as boxplots and averaged curves, respectively. Parametric methods based on MAST, DESeq2, edgeR, and limmatrend showed good $\mathrm{F}_{0.5}$-scores and pAUPRs. Wilcoxon (ranksum) test applied to uncorrected data (denoted as Raw_Wilcox) is currently the most widely used for scRNA-seq DE analysis[23]; however, its performance was relatively low, with or without BEC method.

Next, we checked whether each integrative strategy truly improved the analysis of pooled uncorrected data. ZINB-WaVE provides the "observation weights" (dropout probability) that are used to "unlock" bulk RNA-seq tools to analyze single-cell data[26]. These weights were applied to the parametric methods, edgeR and DESeq2 (denoted as ZW_edgeR and ZW_DESeq2, respectively). First, incorporating the weights of ZINB-WaVE considerably improved edgeR and the use of BEC data from MNNCorrect moderately improved limmatrend (denoted as MNN_limmatrend), in both $\mathrm{F}_{0.5}$-scores and pAUPRs. However, no other BEC data improved DE analysis for these sparse data. Second, most covariate models (tagged with "_Cov") improved their corresponding parametric methods, such as ZW_edgeR, DESeq2, limmatrend, and MAST. In particular, the performance of ZW_edgeR_Cov was among the highest. We also tested four batches and obtained similar results (Supplementary Fig. 1a, c-d).

Model-free simulation tests

We then devised a model-free simulation using real scRNA-seq data to incorporate realistic and complex batch effects and avoid potential bias toward parametric methods. First, we used the two batches from the human pancreas data[27] (named as "human1" and "human2") that were produced by the same laboratory using the same sequencing platform (inDrop[28]). The alpha cells were used for our simulation. Second, we used the two batches from Mouse Cell Atlas (MCA), which were produced by different laboratories using different sequencing platforms (Illumina HiSeq 2500[29] and NovaSeq 6000[30]). For MCA data, the T-cells were used for our simulation. Because these cell types contained several subtypes, the largest clusters that were matched between batches were selected for our simulation. After removing sparsely expressed genes, the overall zero rates of the pancreas and T-cell data were 83% and 73%, respectively. Each batch dataset was randomly split into test and control groups with several different ratios, and then 20% of DE genes (10% up and 10% down) were simulated by downsampling positive counts using binomial distribution (see Methods). As expected, small and large batch effects were observed for the pancreas and MCA data, respectively, after PVCA (Fig. 2a-b). The corresponding $\mathrm{F}_{0.5}$-scores and precision-recall analysis results are shown in Fig. 2c-f. For the pancreas data, the integrative

strategies rarely improved the DE analysis of uncorrected data, and parametric methods, such as limmatrend, DESeq2, and edgeR, performed well with minor differences in pAUPR. However, for MCA data that exhibited large batch effects, several integrative methods were effective. For example, edgeR-based methods exhibited relatively low pAUPRs compared to other parametric methods and the weights of ZINB-WaVE considerably improved edgeR in pAUPR, and incorporating batch covariate further improved the method, making ZW_edgeR_Cov the top-performer in both $\mathrm{F}_{0.5}$-scores and pAUPRs (Fig. 2f). The use of BEC data for ComBat and MNNCorrect also enhanced the performance of limmatrend. However, no improvement was observed for Wilcoxon test with any BEC method.

Whereas BEC methods can be used to reduce batch effects and denoise the data, their benefit in DE analysis was realized for only large batch effects and parametric methods. Covariate modeling also improved their corresponding parametric methods for large batch effects. Many parametric methods outperformed Wilcoxon test, and ZW_edgeR_Cov was among the best performers in both model-based and model-free tests.

Preservation of signs of differential expression

For the simulation cases described above, we counted the DE genes that reversed their signs by using an integrative method to compare the extent of data distortion. For the simulated DE genes, the signs of the computed logFC values in each integrative method were compared with the known ground truth. Meta-analyses were performed in both right- and left-tailed tests and the sign for the smaller *p*-value was used for each gene. Figure 3 and Supplementary Fig. 2 show the ratios of DE genes that altered their signs by each integrative method. Methods based on voom-transformation, including limma, and many meta-analysis methods tended to show relatively high error ratios. Relatively accurate and consistent results were obtained by using Wilcoxon test, MAST, and edgeR-based methods. The performance of Wilcoxon test worsened using BEC methods. We then compared the error ratios among the significantly detected DE genes (q-value < 0.05). In this case, most methods showed a much improved sign prediction. In summary, most integrative methods provided quite accurate results for the analysis of individual genes; however, some of the methods may yield inaccurate results for gene-set level analysis[31,32], as it reflects the expression of all genes.

Effect of sparsity

We also compared the methods for less sparse data. The scRNA-seq data with 60% and 40% zero rates were simulated (see Methods). An overall increase for $\mathrm{F}_{0.5}$-scores and pAUPRs was observed as the data became less sparse (Supplementary Fig. 3). For these data, DESeq2 and limmatrend combined with integrative methods performed well. In particular, the best pAUPRs were achieved using DESeq2 combined with the FEM meta-analysis. Intriguingly, the use of weights from ZINB-WaVE did not improve edgeR anymore for 40% zero rate, whereas the use of BEC data from MNNCorrect consistently improved limmatrend. The pAUPR for Raw_Wilcox ranked only 22nd at the 80% zero rate, but its rank was rather improved to 12th and 15th at the 60% and 40% zero rates, respectively. Whereas all BEC methods tested here did not improve the results of Wilcoxon test at the 80% zero rate, some of the BEC methods improved

the results of this nonparametric method for less sparse data. For example, at the 40% zero rate, Wilcoxon test combined with limma_BEC or ComBat performed better than Raw_Wilcox in both $\mathrm{F}_{0.5}$-scores and pAUPRs, and ranked as high as 5th and 8th in pAUPR, respectively. Overall, the sparsity of data considerably affected the integrative methods. Conventionally used BEC or meta-analysis methods developed for bulk sample data performed well for less sparse data, whereas single-cell-specific strategies such as MAST or ZINB-WaVE achieved a high performance for very sparse data.

Test for heterogeneous samples

Kim and colleagues[33] conducted a comprehensive analysis of scRNA-seq data for LUAD with over 200K cells containing various cell types. Supplementary Fig. 4a shows the distribution of cancer and normal epithelial cells for seven patients with LUAD (stage I). Whereas the clusters of normal cells comprised multiple patients, those of cancer cells were clearly separated between patients. This heterogeneity in cancer samples seems to be related to the different causes or progression of the disease between patients. In other words, different patients may not share exactly the same set of DE genes; this case was referred to as “incomplete association” [12]. We tested this case using four batches with 4% batch-specific DE genes simulated in each batch in addition to 15% common DE genes across the batches. We attempted to detect all the genes that were DE genes in at least one batch (Supplementary Fig. 1b, e-f). The best pAUPR was achieved using MAST_Cov, and ZW_edgeR_Cov also performed well. wFisher that combined the *p*-values of DESeq2 (denoted as DESeq2_wFisher) ranked 2nd in both $\mathrm{F}_{0.5}$-scores and pAUPR. These three methods can be considered when analyzing data complicated with both batch effects and biological variation between samples.

Control of false positives and false discoveries

We used the data for normal epithelial cells in the seven patients with LUAD (stage I) to compare false-positives and false discoveries between integrative methods. The data for each patient were randomly split into two groups with several different ratios (e.g., 2:8, 3:7, 4:6, 5:5), and integrative DE analysis was performed with no DE genes included. We repeated this experiment four times, and the numbers of genes with *p*-value < 0.05 (false-positive) and *q*-value < 0.05 (false discovery) were compared between methods (Fig. 4a). The worst false-positive controls were obtained using the *p*-value combination methods, Fisher and wFisher combined with edgeR or limmatrend. These methods also yielded tens to hundreds of false discoveries. edgeR or edgeR_DetRate showed relatively poor controls of false positives and false discoveries, whereas ZW_edgeR improved the results. Other methods showed a reasonable control of false positives and false discoveries. We obtained similar results for a weaker gene filtering (zero rate < 0.99) (Supplementary Fig. 5).

We then performed the same test for four batches generated by model-based simulation (Fig. 4b). These data represent a simplified condition without subtypes, whereas cell types in real data often include subtypes. Moreover, these data do not represent correlations between genes. However, these two results exhibited some similarity: (1) Poor controls of false-positives and false discoveries were observed using Fisher- and wFisher-based methods. (2) A number of false discoveries were observed using edgeR- and

DESeq2-based methods. (3) Good controls of false-positives and false discoveries were achieved using Wilcoxon test, analysis of pseudobulk data, and single-cell-specific methods such as MAST and ZW_edgeR. We note that an increasing number of false discoveries were observed for increasing variations between samples using Wilcoxon test when “independent” samples were analyzed[23]; however, reliable results were obtained for Wilcoxon test when analyzing “paired” samples that included both test and control conditions in each batch.

Detection of known disease genes

We selected the cells from seven patients with LUAD (stage I), and performed DE analysis between tumor and normal cells for three main cell types: epithelial cells, myeloid cells, and immune cluster composed of T lymphocytes and natural killer cells. These cell types together occupied 68.8% and 74.6% of normal and tumor cells in the scRNA-seq data, respectively (Supplementary Fig. 4b). Because true DE genes are not known for real data, we used the known lung cancer-related genes as the “standard positives”. In total, 221 standard positive genes were obtained from two disease gene databases, DisGeNET[34] and CTD[35]. These genes were weighted by the disease-association score (gda_score > 0.3) provided by DisGeNET (see Methods). All the genes analyzed were sorted by the DE *p*-values in each integrative method, and the cumulative sum of gda_scores of standard positive genes, denoted as cumulative score, was compared between methods in the respective cell types (Fig. 5a-c). In other words, we compared the weighted counts of known disease genes included in the top-*k* DE genes in each method.

To assess the ranks of known disease genes, we devised a truncated Kolmogorov–Smirnov (KS) test that only reflected the ranks of standard positives within the top 20% DE genes, with those in the remaining 80% forced to be evenly distributed. This approach can be particularly useful in selecting methods that are capable of prioritizing important genes in high ranks (see Methods), whereas the conventional KS test risks assessing a large number of middle ranks as significant[36]. Even with this conservative test, many integrative methods exhibited significantly high ranks of the standard positives when epithelial cells were analyzed (*p*-value < 0.01) (Fig. 5d). To compare the performance of integrative DE methods, the area under the cumulative score curves for the top 20% DE genes, denoted as pAUC, was used. The top-performer was ZW_edgeR_Cov, which was closely followed by edgeR_Cov, Raw_Wilcox, limmatrend_Cov, and MAST_Cov (Supplementary Fig. 6a). Covariate modeling marginally improved the corresponding parametric methods, but other integrative strategies hardly improved the DE analyses. This seems to be attributed to the small batch effects between patients (Supplementary Fig. 4c). Interestingly, when myeloid cells and immune cluster were analyzed, none of the integrative methods showed a significance (Fig. 5d).

We then performed DE analysis with the bulk RNA-seq data for LUAD in The Cancer Genome Atlas (TCGA)[37] comprising 493 cancer and 53 normal samples. The corresponding cumulative scores for the known disease genes are also represented in Fig. 5a-c. Remarkably, integrative analysis of epithelial cells for only seven patients outperformed the analysis of hundreds of bulk samples, demonstrating the high potential of scRNA-seq DE analysis to discover disease genes. Figure 5e compares the ranks of 12 genes

with high disease scores (gda_score > 0.5) obtained using five integrative methods and four bulk sample analysis methods. The five integrative methods for scRNA-seq data detected the 12 genes with the average rank percentiles of 33.6% – 40.7% with ZW_edgeR_Cov performing the best, whereas much worse percentiles of 64.2% – 69.7% were achieved using the four TCGA analysis methods. In particular, EGFR, KRAS, CTNNB1, and ERBB2 genes were captured within the top 20% rank by at least three integrative DE methods, and the two genes EGFR and KRAS, which are most common in lung cancer, were ranked in the top 5.3% and 9% by ZW_edgeR_Cov, respectively. In contrast, none of the 12 genes were included within the top 20% by analyses of TCGA data; specifically, EGFR and KRAS were only ranked 61% – 94.7% and 20.4% – 35.8%, respectively. These four genes are known to play important roles in the development of tumor malignancy related to RAS/RAF/MAPK and Wnt signaling pathways[38–40] (see Supplementary Information). The top 20% DE genes in LUAD epithelial cells obtained using three integrative DE methods as well as TCGA analysis results are shown in Supplementary Table 1, which suggests novel LUAD-related genes.

We analyzed two more large-scale bulk sample expression datasets for LUAD that were obtained from GEO database[41] (GSE31210 and GSE43458), where integrative analyses of scRNA-seq data still outperformed the analyses of these bulk sample data in detecting the known disease genes (see Methods and Supplementary Fig. 7a-b). Furthermore, integrative DE methods surpassed both the analyses of scRNA-seq data for individual patients and pseudobulk data[23] (Supplementary Fig. 8). Whereas the analysis of pseudobulk data exhibited strict controls of false positives and false discoveries (Fig. 4), its predictive power for disease-related genes was not high in our analysis of paired data.

Detection of prognostic genes

Next, we performed the same analysis as above using another set of disease-related genes. These genes were selecteded from an integrated survival analysis of five microarray gene expression datasets for patients with LUAD (GSE29013, GSE30129, GSE31210, GSE37745, and GSE50081). The Cox proportional hazards model incorporating covariates of age, sex, and tumor stage[42] was applied to each dataset, and the resulting *p*-values were combined for each gene using wFisher considering the signs of hazards ratios (HRs)[12]. These integrated *p*-values were adjusted for multiple testing correction, yielding 448 genes with *q*-value < 0.05, denoted as “prognostic (standard positive) genes”. We note that only seven of these genes were also included in the 221 known disease genes. Many integrative methods applied to epithelial cells detected the prognostic genes with significantly high ranks, and outperformed both the analyses of TCGA and pseudobulk data (Supplementary Fig. 9). Interestingly, several integrative methods applied to myeloid cells also detected the prognostic genes with significantly high ranks (*p*-value < 0.01), suggesting the association of DE genes in those cell types with the survival of patients.

Pathway analyses for lung epithelial cells and TCGA LUAD data

We further tested the pathway enrichments for scRNA-seq (epithelial cells) and TCGA data to compare the functional relevance of each DE analysis in cancer. The gene-set enrichment analysis (GSEA) was

applied to the ranked genes in each DE method[31,43]. From the pathway database “wikipathway_2021”[44], 192 pathways that were most relevant to cancer progression were selected as standard positives. These pathways were selected on the basis of the ten oncogenic signaling pathways[45] and the seven cancer associated processes[46] as well as those including the keyword(s) tumor, cancer, or carcinoma in their names (see Methods). We classified these pathways into 16 categories for detailed interpretation of the GSEA results (Supplementary Table S2). The cumulative counts of the standard positive pathways showed that GSEA for scRNA-seq data compared favorably with that for TCGA data (Fig. 6).

Interestingly, the analyses of scRNA-seq and TCGA data exhibited distinct functional categories. For example, “Ciliopathies (WP4803)” in the “cell polarity and migration” category ranked first in all the six scRNA-seq analyses, whereas it only ranked 22th to 48th in TCGA analyses. “Genes related to primary cilium development (based on CRISPR) (WP4536)”, which belongs to the same category, was also detected within the top five ranks by all the six scRNA-seq analyses, whereas none of the TCGA analyses detected this pathway. These results represent the cell-type-specific perturbation of pathways in lung epithelial cells. Several pathways in the “cell survival” category, including “Apoptosis (WP254)” and “Senescence and Autophagy in Cancer (WP615)”, were also detected in scRNA-seq analyses, whereas none of them in that category was detected in TCGA analyses. Moreover, the five categories “WNT”, “PI3K”, “HIPPO”, “NOTCH”, and “P53” in oncogenic signalling pathways were detected in scRNA-seq analyses, but none of them were detected in TCGA analyses. In contrast, GSEA for TCGA data detected five and seven pathways in the two categories “genomic instability” and “inflammation”, respectively, whereas GSEA for scRNA-seq data detected none and at most three in the respective categories. By analyzing the epithelial cell data, we were able to detect many canonical oncogenic pathways as well as cell-type-specific pathways that the bulk sample analyses missed. The GSEA results for selected integrative DE methods and TCGA analysis results are available from Supplementary Table 3.

A gross comparison

All the test results in this study, including the speed and scalability of integrative methods are summarized in Fig. 7. pAUPR shows how precisely a method can detect DE genes while maintaining a low type I error and the results are summarized in Fig. 7a. Other measurements are shown in Fig. 7b. We have classified the methods into three levels (“high”, “medium” and “low”) in each category based on the test results for all datasets used in this study: model-based simulation for three sparsity levels (40%, 60%, and 80%), four-batch cases with or without additional biological variations, three model-free simulation cases, and three cell types in the lung cancer data. See Methods for the criteria used in each category.

Discussion

Here, we benchmarked various DE methods for scRNA-seq data with multiple batches. We found that the use of BEC data rarely improved DE analysis for sparse scRNA-seq data. One exception was MNNCorrect that enhanced limmatrend for large batch effects. The use of BEC data from ZINB-WaVE did not improve DE analysis; however, the observation weights of ZINB-WaVE considerably improved the bulk RNA-seq

tool edgeR, and incorporating batch covariate further improved the performance (ZW_edgeR_Cov). Covariate modeling overall improved the performance of the corresponding parametric DE methods, when substantial batch effects were involved. In contrast, the benefit of integrative strategies was not clearly observed for data exhibiting minor batch effects, as demonstrated in pancreas and LUAD scRNA-seq data analyses. Therefore, the several strategies suggested in this study (Fig. 7) are particularly useful for scRNA-seq DE analysis across different sequencing platforms or laboratories.

Batch effects also impacted the widely used Wilcoxon test. For small batch effects, many DE methods showed similarly good performance, and the pAUPR of Wilcoxon test was only 1.7% point smaller than that of the top-performer. However, for large batch effects, its pAUPR was 12.2% point smaller than that of the top-performer (Fig. 2e, f). This indicates that the simple Wilcoxon test without any data correction is a reasonable choice for scRNA-seq data with small batch effects. Indeed, Wilcoxon test performed well in detecting the disease genes from LUAD scRNA-seq data (Supplementary Fig. 6a). The sparsity of data also affected DE analysis. For less sparse data (zero rate = 60% or 40%), the conventionally used BEC (e.g., ComBat) and meta-analysis methods showed a good performance, whereas single-cell specific methods, such as ZINB-WaVE weights and MAST, became less advantageous.

Our results suggest using parametric methods with batch covariate over Wilcoxon test for scRNA-seq DE analysis, especially for data exhibiting substantial batch effects. In particular, edgeR/edgeR_Cov incorporating the weights of ZINB-WaVE and MAST/MAST_Cov exhibited notably consistent and good performances in various tests, except the long computation time of ZINB-WaVE. For example, it took approximately one hour using ZW_edgeR_Cov to process the two batches of MCA T-cell data with 3059 genes and 625 cells in total; however, it took as long as 19 hours for the seven batches of LUAD data with 10,238 genes and 7,746 cells in total, using Intel Core i7-8700 CPU (3.2 GHz).

We note that not all the integrative strategies improved the DE analysis of sparse scRNA-seq data. For example, most BEC methods deteriorated the result of Wilcoxon test, and methods that used voom-transformed data and meta-analysis methods exhibited relatively high error ratios in the sign prediction of DE genes. In addition, the *p*-value combination methods, Fisher/wFisher, tended to exhibit a poor control of false-positives. Thus, these methods are not recommended for integrative DE analysis of sparse scRNA-seq data.

Finally, we tested whether integrative scRNA-seq DE analysis could be used to prioritize disease-related genes better than analyses of bulk or pseudobulk data. This was verified for two independently derived sets of disease genes and three large-scale bulk sample datasets for LUAD. For a relevant cell type, many integrative scRNA-seq DE methods exhibited a superior predictive power for disease-related genes compared to the analyses of bulk sample data, and pathway analyses further supported this finding. Overall, our results suggest using integrative analysis of scRNA-seq data considering cells as independent replicates rather than using bulk or pseudobulk data to discover disease genes. For scRNA-seq data with large batch effects, several high-performing integrative methods have been suggested.

Methods

F_β -score and partial area under precision-recall curve

In DE analysis of scRNA-seq data, it is often important to identify a small number of genes (markers) that are capable of characterizing each cell type. Moreover, it is not reasonable to expect to identify all DE genes from highly noisy and sparse data. Thus, we use generalized F-score (F_β) and partial AUPR (pAUPR) that weigh precision twice higher than recall to assess a DE analysis method. In binary classification task, F-score is the harmonic mean of precision and recall. For a list of DE genes (*q*-value < 0.05), we use F_β ($\beta = 0.5$) defined as follows:

$$F_\beta = \frac{(1+\beta^2) \bullet precision \bullet recall}{\beta^2 \bullet precision + recall}, \beta > 0$$

The F_β-scores were calculated for both up and downregulated genes and both results were included in Fig. 1 and Fig. 2. To assess general performance of a classifier, precision-recall curve has often been considered. Instead of using the whole AUPR, we suggested using pAUPR (T = 0.5) defined as follows:

$$pAUPR_T = \frac{1}{T}\int_0^T precision_t dt, 0 < T < 1$$

Model-based simulation for 80%, 60%, and 40% zero rates

Splatter R package[24] was used to simulate scRNA-seq data based on negative binomial model. The dropout parameter values *dropout.mid* = 0.01–0.05, 1.1–1.3, and 3.7–3.8 were used to simulate data with overall zero rates 40%, 60%, and 80%, respectively. *splatSimulate* function was used to simulate different batches. Large batch effects (*batch.facLoc* = 0.4 and *batch.facScale* = 0.4) and small group differences (*de.facLoc* = 0.2 and *de.facScale* = 0.2) were simulated to clearly demonstrate the effects of BEC. We created six scenarios for combinations of two dropout values and three group size ratios (2:8, 3:7, 4:6) for each sparsity level. The batch sizes with 300 and 750 cells were used for the two-batch case; and 300, 400, 400, and 750 were used for the four-batch case. Approximately 2,500 genes survived gene filtering and included 800–1,000 DE genes (half up and half downregulated) for the parameters given above. A smaller proportion of DE genes was tested in our “model-free” simulation. Then, a more complicated case was tested. First, four batch groups with small batch effects (*batch.facLoc* = 0.0 and *batch.facScale* = 0.1) were simulated. We simulated 15% DE genes common to all batches and 4% batch-specific DE genes by downsampling their read counts in the test condition (see **Model-free simulation** section below). Thereby, four control groups with relatively similar distributions and four test groups with more heterogeneous distributions were obtained (Supplementary Fig. 2b). This test was designed to imitate the heterogeneity of samples in complex disease such as cancer where each patient may exhibit different progression or cause of disease.

Model-free simulation

MCA and pancreas data were used to simulate scRNA-seq data. MCA data comprised two independent datasets obtained using different sequencing techniques. The original data included two batches containing 4,239 and 2,715 cells. We chose T-cells for our simulation. Because T-cells included several subtypes, we selected the largest clusters from each batch that shared marker genes identified by "FindMarkers" function in Seurat package[16]. Specifically, we selected clusters with 358 and 266 T-cells from different batches. For pancreas data, we used the clusters with 241 and 659 alpha-cells from "human1" and "human2" batches, respectively. Then, each batch dataset was randomly divided into test and control groups with different ratios to cover several scenarios. We then randomly selected two groups of genes, each with 10% of all genes; one group was downsampled in the test group and the other downsampled in the control group using binomial distribution to simulate DE genes. The success probability for the binomial distribution was sampled from the beta distribution with the shape parameters $\alpha = \beta = 2$ that are expected to generate DE genes with the median fold-change two.

Covariate modeling

The log-linear model[10,47] has been frequently used to model the read count data as follows:

$$\log\left(\mathrm{E}\left(y_{ij}\right)\right) = \alpha_{i0} + \alpha_{i1} L_j + \sum_{b=1}^{B} \beta_{jb} I_{jb} + \sum_{g=1}^{G} \gamma_{jg} I_{jg}$$

where y_{ij} is the read count of gene *i* and sample *j*, L_j is the library size of sample *j*, *I* is the indicator function a specific sample group, α's, β's, and γ's are the model parameters, *B* and *G* are the numbers of batches and sample groups used, respectively. Then, DE of a gene can be tested by comparing models with or without the group labels (likelihood ratio). It is expected that incorporating the batch covariate improves DE analysis across different batches.

Principal variance component analysis

Principal variance component analysis (PVCA)[25] was used to estimate the variability of experimental effects. PVCA combines principal component analysis (PCA) and variance components analysis (VCA) to take the advantages from both techniques. PCA reduces the dimension of data while preserving their major variability. VCA fits a mixed linear model using the factors of interest to estimate and partition the total variability. Whereas PVCA is a generic approach used to quantify the proportion of variations of different effects, it provides handy assessment for the batch effects before and after the correction. Supplementary Fig. 1 shows PVCA results for six BEC methods applied to MCA T-cell data and corresponding tSNE plots[48] before and after the correction.

Collection of known disease genes

Two disease gene databases, DisGeNET and CTD were used to retrieve known lung cancer genes. In DisGeNET, 2438 genes were annotated with term, "Adenocarcinoma of lung (disorder)". DisGeNET provides gene-disease association score (gda_score), which is weighted sum of the number of each level/type of sources, and the number of publications supporting the association. Among the 2438 genes,

we have selected only 207 genes with gda_score 0.3 or larger. In CTD, we have selected 28 genes that were annotated with “Adenocarcinoma of Lung” and curated as “Marker/mechanism” in “Direct.Evidence” field. Among them, 14 genes that were also selected from DisGeNET and their median score was given to the rest 14 genes that were exclusively selected from CTD. In total, 221 genes were used as standard positives.

Categorization of standard positive pathways in lung cancer

The standard positive pathways were categorized on the basis of ten oncogenic signaling pathways[45] and seven cancer associated processes[46]. The ten oncogenic signaling pathways included (1) cell cycle, (2) Hippo signaling, (3) Myc signaling, (4) Notch signaling, (5) oxidative stress response/Nrf2, (6) PI-3-Kinase signaling, (7) receptor-tyrosine kinase (RTK)/RAS/MAP-Kinase signaling, (8) TGFβ signaling, (9) p53, and (10) β-catenin/Wnt signaling. Here, “Myc signaling” category was not detected by any analysis, so was excluded. The seven cancer associated processes included (1) cell proliferation, (2) cell polarity and migration, (3) cell survival, (4) cell metabolism, (5) cell fate and differentiation, (6) genomic instability, and (7) tumor microenvironment. Among them, “cell cycle”, "cell proliferation” and “cell fate and differentiation” were combined into one category, and “tumor microenvironment” was divided into its subcategories, “inflammation” and “angiogenesis”. Lastly, pathways that included the keywords, tumor/cancer/carcinoma in their names were collected into a separate category, where less relevant pathways such as retinoblastoma or glioblastoma were excluded. In total, 190 standard positive pathways for cancer were classified into 16 categories (Supplementary Table 3).

Analyses of LUAD bulk-sample expression data

DE analysis for TCGA LUAD bulk RNA-sequencing data between 493 cancer and 53 normal samples were performed incorporating covariates age, sex, and smoking history using four methods, DESeq2, edgeR, limma and limmatrend methods. 19201 genes with five or larger mean count that were commonly found in filtered epithelial scRNA-seq data were analyzed. Two LUAD microarray expression datasets (GSE31210 and GSE43458) were also analyzed. The former consisted of 226 tumor and 15 normal samples and covariates of age, sex, and smoking history were incorporated in DE analysis. The data were normalized by MAS5 and the log-transformed data were used for limmatrend. The latter consisted of 80 cancer and 30 normal samples. Only smoking history was available and used as covariate. RMA normalization and limmatrend were used.

Criteria for classifying performance

Speed

We used LUAD epithelial cell data for the seven patients to compare the computing times between integrative DE methods and classified them based on their ranks as follows

- High: the top 20% fastest
- Medium: between “High” and “Low”

- Low: bottom 20%

Scalability

We compared the proportionality between the computing time and the data size. We estimated this coefficient for the square root of the number of data entries (cells × genes). For dataset i including N_i cells and M_i genes, the computing time T_i (seconds) of method K was modeled as

$$T_i = \alpha_K \sqrt{N_i \bullet M_i}$$

The scalability of method K was classified based on the coefficient α_K as follows:

- High: $\boldsymbol{\alpha_K}$ < 2
- Medium: between “High” and “Low”
- Low: $\boldsymbol{\alpha_K}$ > 4

Sign preservation

Because the ranges of values are different between the results of datasets, we think of aggerating the relative difference between boxplots/groups of performance values of methods. The percentage of errors (P) is calculated based on the difference between medians (DBM) and the overall visible spread (OVS) as

$$\mathrm{P} = \frac{\mathrm{DBM}}{OVS} \times 100$$

We used K-means algorithm to cluster the error ratios for four simulationdatasets to into three groups.

pAUPR, AUPR, -score, F-score

These scores are classified based on the estimated ranks across nine simulation datasets.

- High: if the method surpasses Raw_Wilcox in at least six out of the nine datasets.
- Medium: between “High” and “Low”
- Low: if the method surpasses Raw_Wilcox in at most three out of the nine datasets.

Truncated Kolmogorov-Smirnov test

Kolmogorov-Smirnov (KS) test assesses the maximum distance between empirical and null cumulative distribution functions (cdf). The empirical distribution was generated by accumulating the gene scores of standard positives in the order of DE gene p-values and the test statistic is given as follows:

$$F_x(u) = \frac{\sum_{i=1}^{u} w_i}{\sum w_i}, cdfofempiricaldistribution$$

$$F_y(u) = cdfofnullhypothesis$$

KS statistic (right tailed):

KS statistic (right tailed): $\boldsymbol{D^+ = max\left(F_x(u) - F_y(u)\right)}$ where $\boldsymbol{w_i}$'s are the weights of standard positive genes.

If the $\boldsymbol{ith}$ gene does not belong to standard positive genes, $\boldsymbol{w_i = 0}$.

A drawback of KS test is that the maximum discrepancy $\boldsymbol{D^+}$ can occur for a low gene rank[36]. Because we are interested in methods that are capable of prioritizing standard positive genes in high ranks, we modified the statistic so that $\boldsymbol{D^+}$ can occur only within top 20% ranks as follows:

wKS statistic (right tailed): $\boldsymbol{\tilde{D}^+ = max\left(\widetilde{F_x}(u) - F_y(u)\right)}$

$$\tilde{F}_x(u) = \begin{cases} F_x(u), & u < N \\ F_x(N) + \frac{(u-N)}{(u_{max}-N)} \bullet (1 - F_x(N)), & u \geq N \end{cases}$$

where $\mathbf{u_{max}}$ is total number of genes $\mathbf{N}$ corresponds to the top 20% rank. In other words, the ranks of standard positives out of the top 20% were "uniformized" not to affect the test result.

Declarations

Competing interests

The authors declare no competing interests.

Additional Information

Availability of data and materials

The code used for our simulation tests is available from Github (https://github.com/noobCoding/Benchmarking-integration-of-scRNAseq-differential-analysis).

References

1. Park, J. *et al.* Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. Science **360**, 758–763, doi:10.1126/science.aar2131 (2018).
2. Lambrechts, D. *et al.* Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med **24**, 1277–1289, doi:10.1038/s41591-018-0096-5 (2018).

3. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol **21**, 12 (2020).
4. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. Nat Methods **19**, 41–50, doi:10.1038/s41592-021-01336-8 (2022).
5. McDavid, A. *et al.* Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinformatics **29**, 461–467, doi:10.1093/bioinformatics/bts714 (2013).
6. Argelaguet, R., Cuomo, A. S. E., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. Nat Biotechnol **39**, 1202–1215, doi:10.1038/s41587-021-00895-7 (2021).
7. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol **15**, e8746, doi:10.1186/s13059-019-1850-9 (2019).
8. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England) **26**, 139–140, doi:10.1093/bioinformatics/btp616 (2010).
9. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).
10. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res **43**, e47, doi:10.1093/nar/gkv007 (2015).
11. Wang, X. *et al.* An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. Bioinformatics **28**, 2534–2536, doi:10.1093/bioinformatics/bts485 (2012).
12. Yoon, S., Baik, B., Park, T. & Nam, D. Powerful p-value combination methods to detect incomplete association. Sci Rep **11**, 6980, doi:10.1038/s41598-021-86465-y (2021).
13. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. Nature Communications **9**, 284, doi:10.1038/s41467-017-02554-5 (2018).
14. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nature Biotechnology **36**, 421–427, doi:10.1038/nbt.4091 (2018).
15. Lin, Y. *et al.* scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proceedings of the National Academy of Sciences* **116**, 9775, doi:10.1073/pnas.1820006116 (2019).
16. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. Cell **177**, 1888–1902.e1821, doi:https://doi.org/10.1016/j.cell.2019.05.031 (2019).
17. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics **8**, 118–127, doi:10.1093/biostatistics/kxj037 (2006).
18. Lun, A. T., Chen, Y. & Smyth, G. K. It's DE-licious: A Recipe for Differential Expression Analyses of RNA-seq Experiments Using Quasi-Likelihood Methods in edgeR. Methods Mol Biol **1418**, 391–416,

doi:10.1007/978-1-4939-3578-9_19 (2016).

19. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. Nature Methods **15**, 255–261, doi:10.1038/nmeth.4612 (2018).
20. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology **15**, R29, doi:10.1186/gb-2014-15-2-r29 (2014).
21. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol **3**, Article3, doi:10.2202/1544-6115.1027 (2004).
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. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. Nature Communications **12**, 5692, doi:10.1038/s41467-021-25960-2 (2021).
24. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol **18**, 174, doi:10.1186/s13059-017-1305-0 (2017).
25. Li, J., Bushel, P. R., Chu, T.-M. & Wolfinger, R. D. in Batch Effects and Noise in Microarray Experiments 141–154 (2009).
26. Van den Berge, K. *et al.* Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. Genome Biol **19**, 24, doi:10.1186/s13059-018-1406-4 (2018).
27. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Systems **3**, 346-+ (2016).
28. Klein, Allon M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell **161**, 1187–1201, doi:https://doi.org/10.1016/j.cell.2015.04.044 (2015).
29. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. Cell **172**, 1091–1107.e1017, doi:10.1016/j.cell.2018.02.001 (2018).
30. Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature **562**, 367–372, doi:10.1038/s41586-018-0590-4 (2018).
31. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550, doi:10.1073/pnas.0506580102 (2005).
32. Nam, D. & Kim, S. Y. Gene-set approach for expression pattern analysis. Brief Bioinform **9**, 189–197, doi:10.1093/bib/bbn001 (2008).
33. Kim, N. *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat Commun **11**, 2285, doi:10.1038/s41467-020-16164-1 (2020).
34. Pinero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic acids research **48**, D845-D855, doi:10.1093/nar/gkz1021 (2020).
35. Davis, A. P. *et al.* Comparative Toxicogenomics Database (CTD): update 2021. Nucleic Acids Res **49**, D1138-D1143, doi:10.1093/nar/gkaa891 (2021).

36. Damian, D. & Gorfine, M. Statistical concerns about the GSEA procedure. *Nat Genet* **36**, 663; author reply 663, doi:10.1038/ng0704-663a (2004).
37. Collisson, E. A. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. Nature **511**, 543–550, doi:10.1038/nature13385 (2014).
38. Sergina, N. V. & Moasser, M. M. The HER family and cancer: emerging molecular mechanisms and therapeutic targets. Trends Mol Med **13**, 527–534, doi:10.1016/j.molmed.2007.10.002 (2007).
39. van Roy, F. & Berx, G. The cell-cell adhesion molecule E-cadherin. Cell Mol Life Sci **65**, 3756–3788, doi:10.1007/s00018-008-8281-1 (2008).
40. Yang, H., Liang, S. Q., Schmid, R. A. & Peng, R. W. New Horizons in KRAS-Mutant Lung Cancer: Dawn After Darkness. Front Oncol **9**, 953, doi:10.3389/fonc.2019.00953 (2019).
41. Clough, E. & Barrett, T. The Gene Expression Omnibus Database. Methods Mol Biol **1418**, 93–110, doi:10.1007/978-1-4939-3578-9_5 (2016).
42. Cox, D. R. Regression Models and Life-Tables. J R Stat Soc B **34**, 187-+ (1972).
43. Sergushichev, A. A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. bioRxiv, 060012, doi:10.1101/060012 (2016).
44. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. Curr Protoc **1**, e90, doi:10.1002/cpz1.90 (2021).
45. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337 e310, doi:10.1016/j.cell.2018.03.035 (2018).
46. Sever, R. & Brugge, J. S. Signal transduction in cancer. Cold Spring Harb Perspect Med **5**, doi:10.1101/cshperspect.a006098 (2015).
47. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol **20**, 296, doi:10.1186/s13059-019-1874-1 (2019).
48. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. Journal of Machine Learning Research **9**, 2579–2605 (2008).

Figures

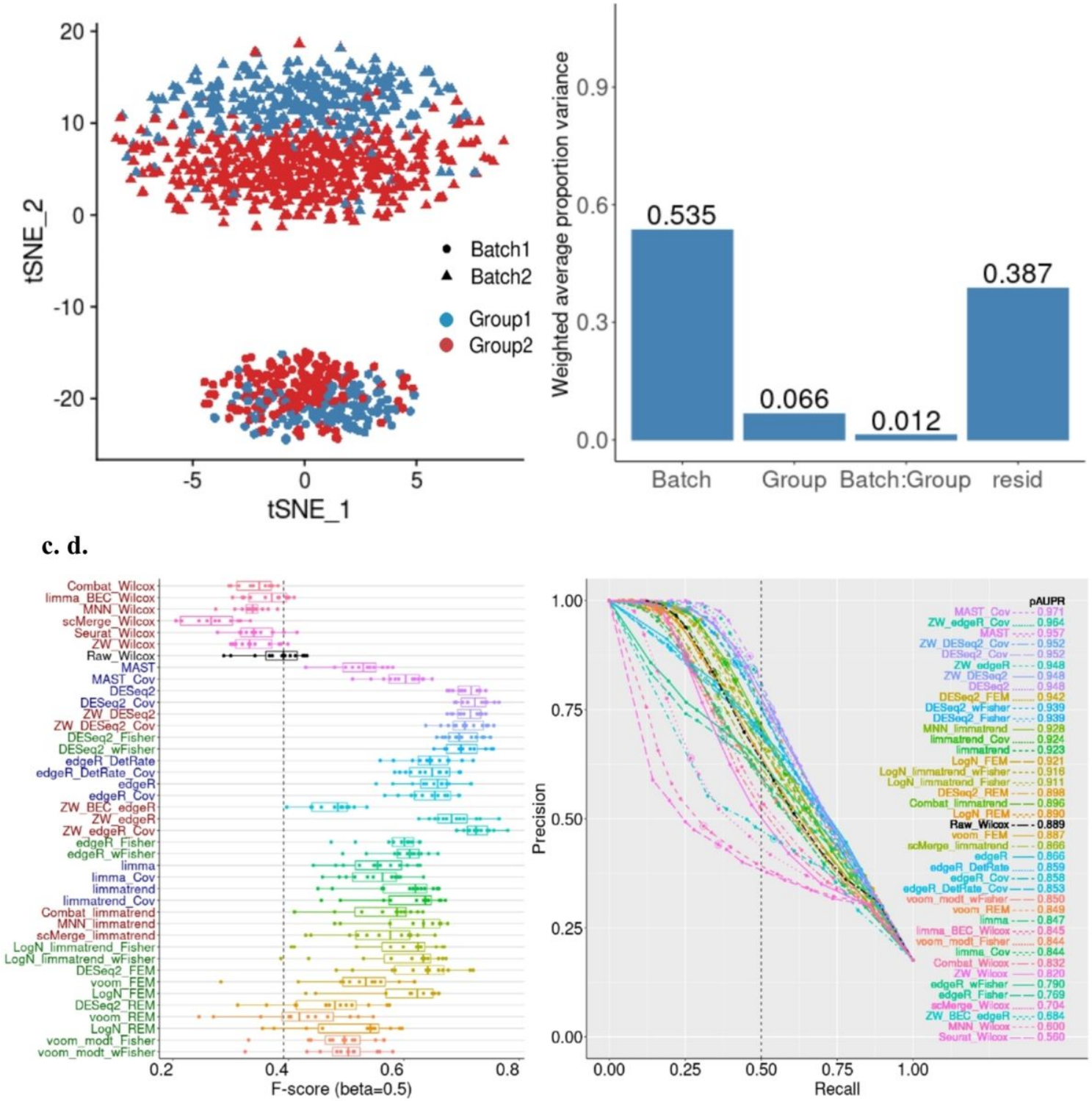

Figure 1

Model-based simulation results. a Scatter plot of the simulated data. Two t-SNE components are represented. **b** PVCA results, representing large batch effects and small group difference. **c** $F_{0.5}$-scores for 41 integrative DE methods. Results for several different scenarios are represented as boxplots. **d** Precision-recall curves. The partial areas under the curve for recall rate < 0.5 (pAUPRs) are computed and sorted in descending order in the legends.

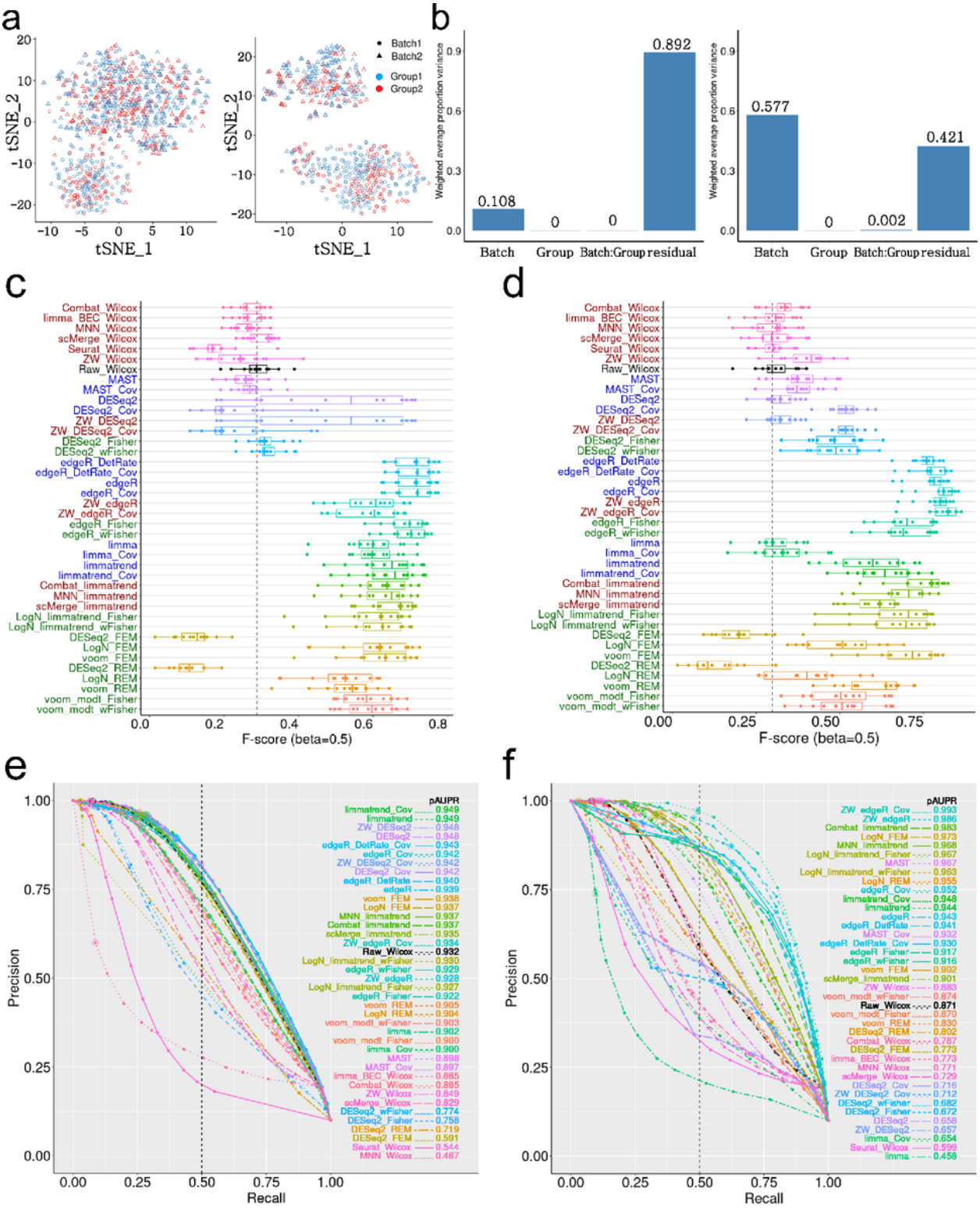

Figure 2

Model-free simulation results. **a** Scatter plots for pancreas (left) and MCA (right) data. **b** PVCA results for pancreas (left) and MCA (right) data, representing small and large batch effects, respectively. The

-scores for **c** pancreas and **d** MCA data. Precision-recall curves for the 41 methods for **e** pancreas and **f** MCA T-cell data. The partial areas under the curve for recall rate < 0.5 (pAUPR) are computed and sorted

in descending order in the legends.

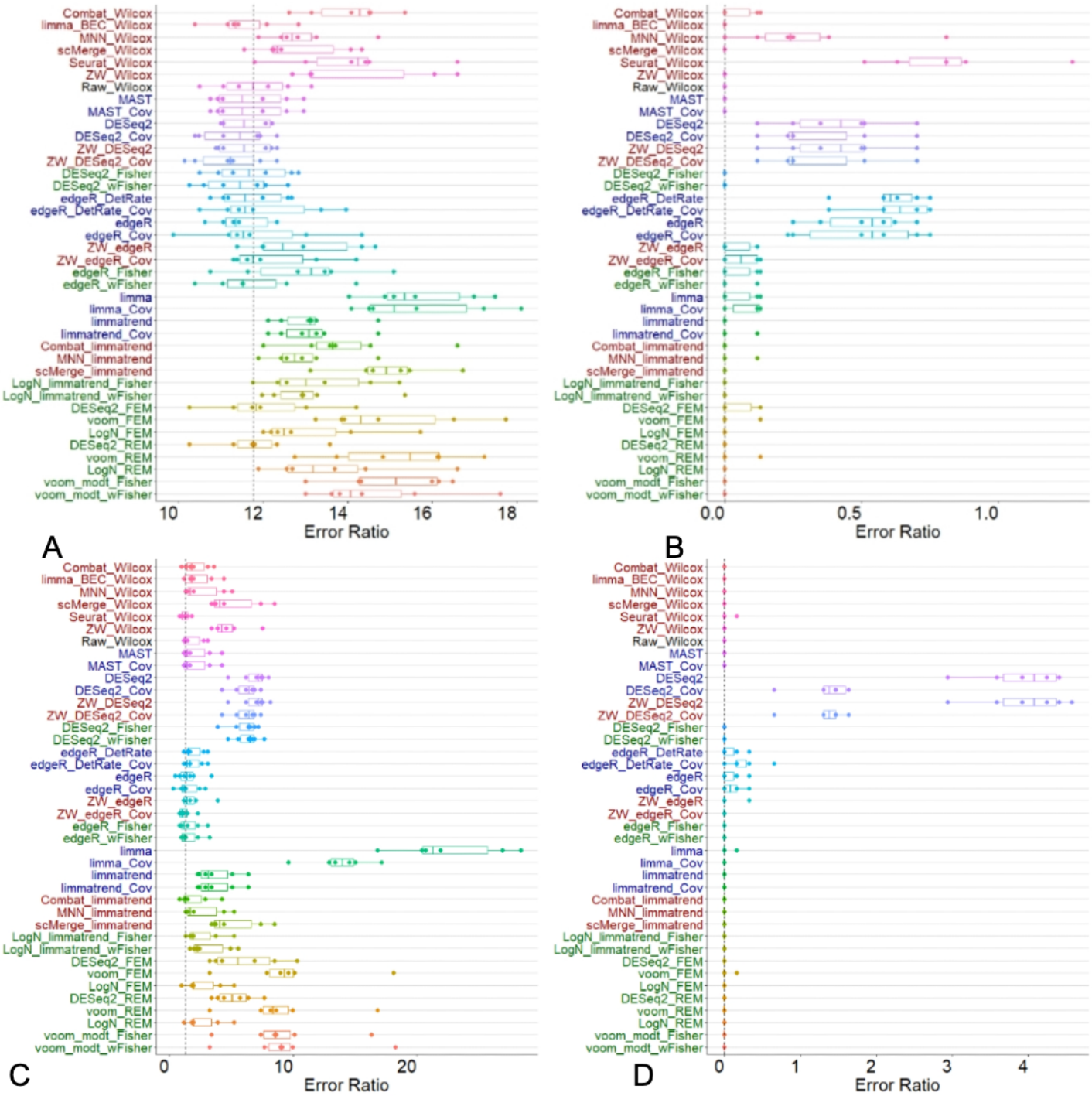

Figure 3

Ratios of DE genes that altered their signs by using integrative DE methods. **a** Total error ratios for model-based simulation data (two batches). **b** Error ratios for model-based simulation data among the significant genes. **c** Total error ratios for model-free simulation data (MCA T-cell). **d** Error ratios for model-free simulation data (MCA T-cell) among the significant genes.

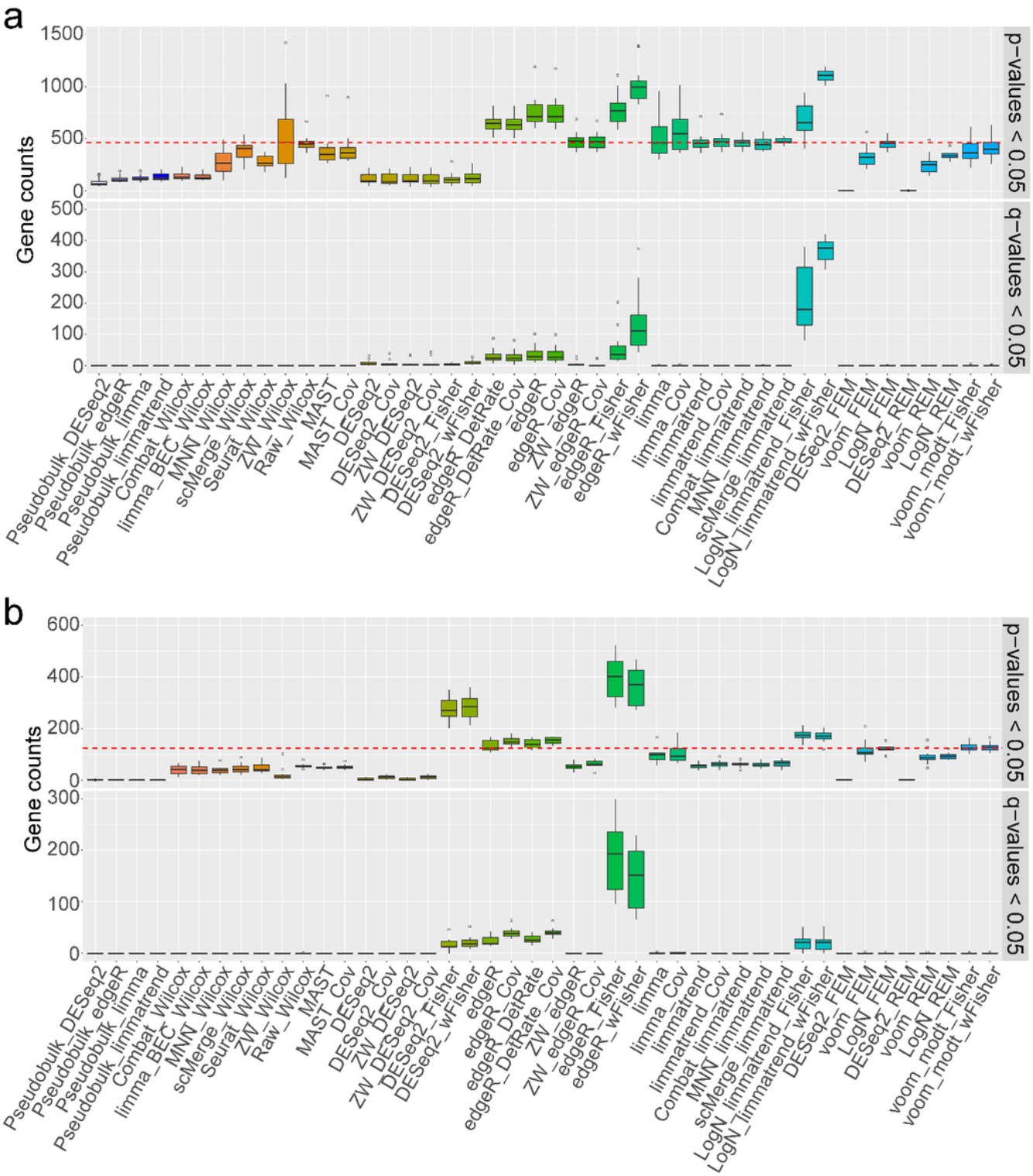

Figure 4

Comparison of false positive controls (p-value < 0.05) and false discovery controls (q-value < 0.05) between integrative DE and pseudobulk analysis methods (gene filtering: zero rate < 0.95). Test results for a seven batches of normal lung epithelial cells and b four batches of model-based simulation data. Red dashes indicate the five percent of all genes tested.

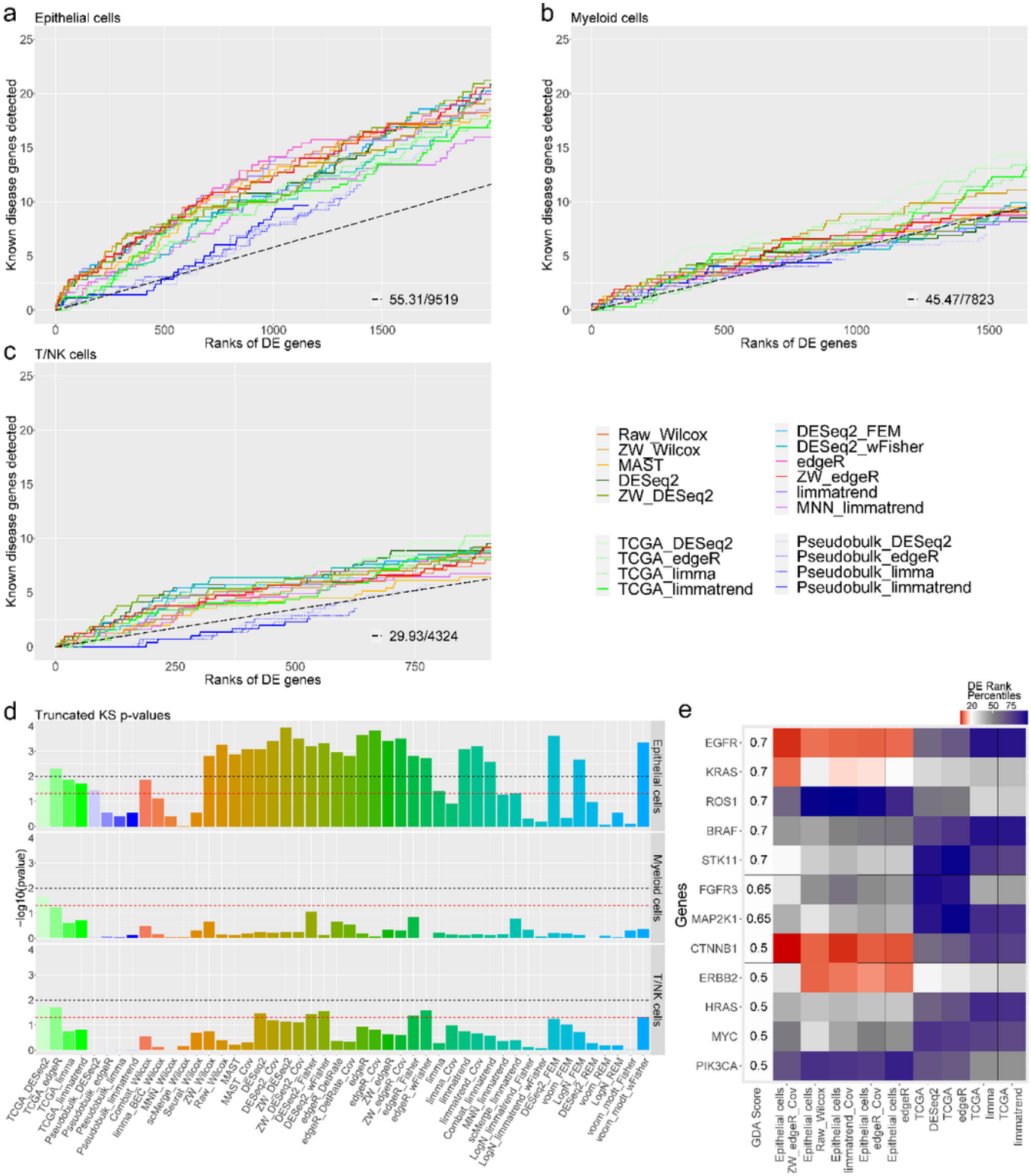

Figure 5

Detection of known LUAD genes is compared between DE analyses of scRNA-seq, TCGA LUAD RNA sequencing, and pseudobulk data. Cumulative gda_scores up to top 20% DE genes are shown for three cell types: **a** epithelial cells, **b** myeloid cells and **c** T/NK cells. X-axis represents the gene ranks in each DE analysis (up to top 20%). Y-axis represents the cumulative score of of known disease genes captured within top-*k* gene ranks by each DE method. The black-dashed slopes represent the expected cumulative

scores of known disease genes for random gene ranks. 15 and 4 methods are selected for analysis of scRNA-seq and bulk/pseudobulk data, respectively. **d** *p*-values for the scRNA-seq DE methods and the bulk/pseudobulk analysis methods (truncated KS test) are shown for the three cell types. Black and red dashes represent the two significance cutoffs *p*-values < 0.01 and < 0.05, respectively. **e** Rank percentages of the 12 known LUAD genes with gda_score no less than 0.5 are visualized for five integrative DE methods (epithelial cells) and four bulk sample (TCGA) DE methods.

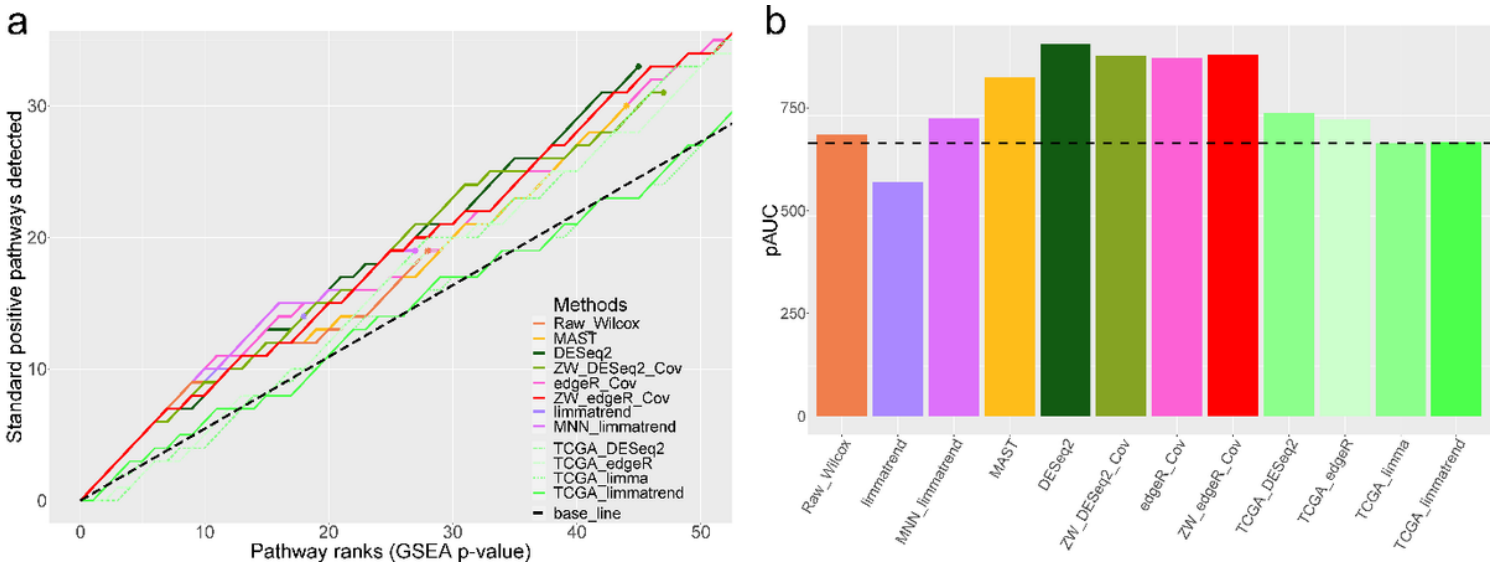

Figure 6

Comparison of tumor associated pathways detected using GSEA from DE analysis of scRNA-seq and TCGA data. a Cumulative counts of tumor associated pathways detected from DE analysis of LUAD epithelial cells and TCGA data are compared (up to top 50 significant pathways). X-axis represents the pathway ranks in each DE analysis. The dashed slope (black) represents the expected numbers of tumor associated pathways for random pathway ranks. Eight and four methods are selected to analyze scRNA-seq and TCGA data, respectively. **b** pAUCs for the top 50 pathways corresponding to each GSEA result are compared. Black-dashed line represents the expected pAUC for the random pathway ranks.

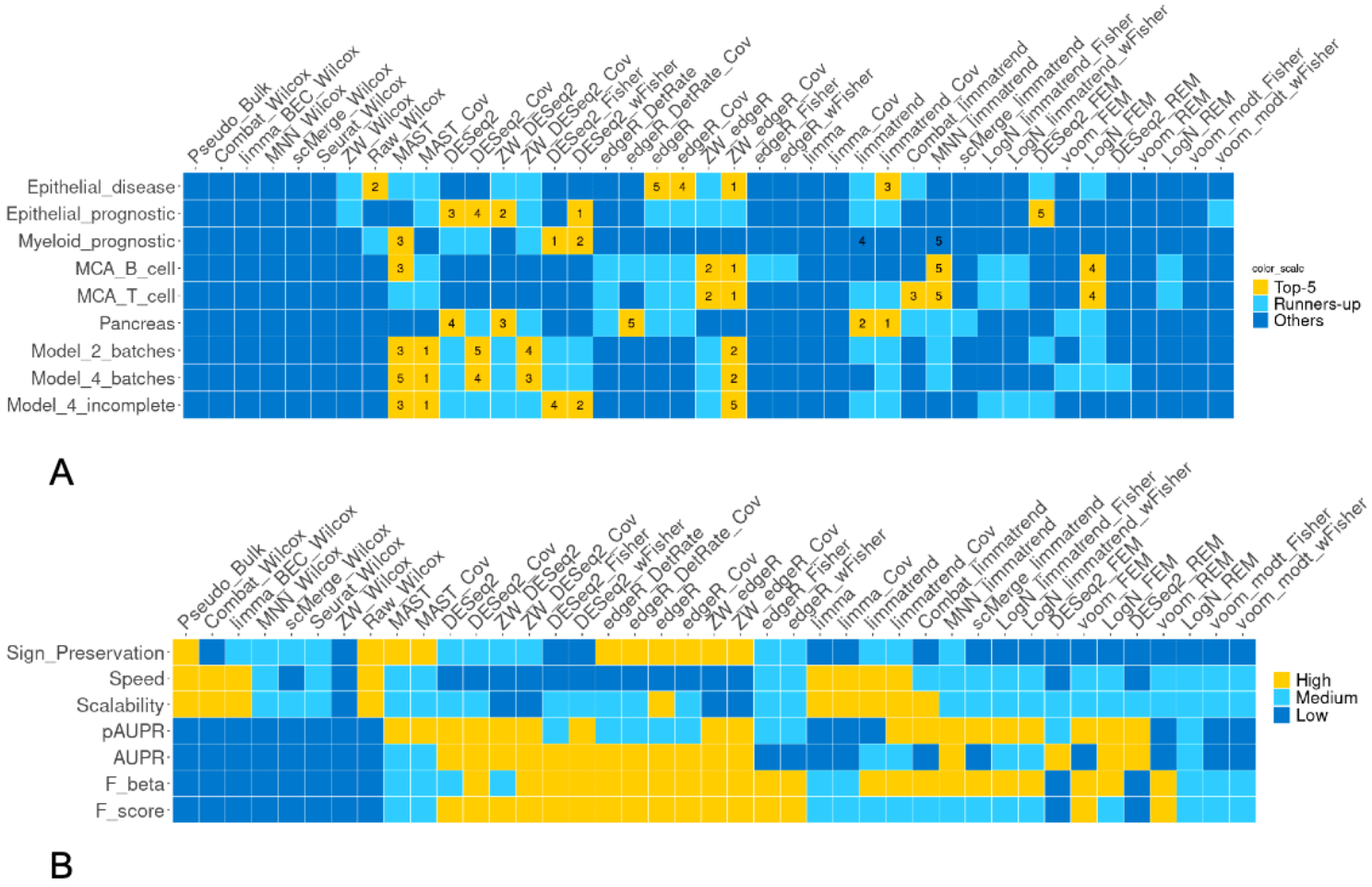

Figure 7

Performance classification of DE methods. a DE methods are classified into the top-5, the next ten ‘Runners-up’, and the remaining on the basis of pAUC for LUAD scRNA-seq data (up to top 20% gene ranks) and pAUPR for simulated scRNA-seq data (recall rate < 0.5). This summary provides an overview on how precisely each DE method detects the true DE genes across simulation and real datasets. **b** Summaries of several performance measures across simulation datasets except for “Speed”. Speeds of each method are measured using the epithelial cell data for the seven patients with LUAD.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 3722870supp6601037rcwsrc.docx
- supplementarytables.zip