OXFORD

A comprehensive assessment of cell type-specific differential expression methods in bulk data

Guanqun Meng, Wen Tang, Emina Huang, Ziyi Li and Hao Feng

Corresponding author. Hao Feng, Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH 44106, USA,
E-mail: hxf155@case.edu

Abstract

Accounting for cell type compositions has been very successful at analyzing high-throughput data from heterogeneous tissues. Differential gene expression analysis at cell type level is becoming increasingly popular, yielding biomarker discovery in a finer granularity within a particular cell type. Although several computational methods have been developed to identify cell type-specific differentially expressed genes (csDEG) from RNA-seq data, a systematic evaluation is yet to be performed. Here, we thoroughly benchmark six recently published methods: CellDMC, CARseq, TOAST, LRCDE, CeDAR and TCA, together with two classical methods, csSAM and DESeq2, for a comprehensive comparison. We aim to systematically evaluate the performance of popular csDEG detection methods and provide guidance to researchers. In simulation studies, we benchmark available methods under various scenarios of baseline expression levels, sample sizes, cell type compositions, expression level alterations, technical noises and biological dispersions. Real data analyses of three large datasets on inflammatory bowel disease, lung cancer and autism provide evaluation in both the gene level and the pathway level. We find that csDEG calling is strongly affected by effect size, baseline expression level and cell type compositions. Results imply that csDEG discovery is a challenging task itself, with room to improvements on handling low signal-to-noise ratio and low expression genes.

Keywords: deconvolution, RNA-seq, heterogeneous samples, cell type-specific signal, differentially expressed genes

Introduction

Identifying differentially expressed genes (DEG) in bulk RNA-seq data is often the primary task in clinical studies where researchers want to find gene expression alterations associated with phenotypes of interest, such as neurodegenerative and neuropsychiatric disorders [1–7], virus infection [8–10], inflammation [11, 12], aging [13, 14], smoking history [15, 16] and many more. In these studies, the real clinical samples for sequencing often contain a mixture of different cell types. For example, blood samples have a mixture of B cells, memory T cells, CD4 T helper cells, natural killer cells, dendritic cells and others [17]. The real clinical RNA-seq signal is, in fact, a mosaic of at least several pure cell types. The bulk RNA-seq signals, as a result, are the weighted average of signals from multiple pure cell types. The weights, naturally, are the proportions of the associated cell types.

With the cell types mixture process described above, it is easy to identify two potential causes that yield DE genes in observed bulk data: cell composition change or cell type expression level change. For example, gene expression alterations in lung cancer before and after immunotherapy do not necessarily indicate the tumor's

transcriptome change; instead, it reflects an increased content of tumor-infiltrating lymphocytes [18]. As another example, the proportion of cytotoxic CD8+ T cells would decrease in late-stage melanoma samples [19]. Therefore, in Differential Expression (DE) analysis, the mixing proportions can confound the phenotype of interest. Traditional DE detection methods, in general, fail to consider the cell mixture problem. Properly accounting for cell type mixture in DE gene detection became an active research domain in the last several years.

Several computational methods were recently developed to take sample mixture proportions into modeling, to identify cell type-specific DEG (csDEG). Relevant methods include CARseq [20], TOAST [21], CeDAR [22], CellDMC [23], LRCDE [24], TCA [25], csSAM [26], HIRE [27] and DESeq2 [28]. A schematic overview is shown in Figure 1. In general, these methods can take bulk RNA-seq data as the input, together with cell type mixing proportions given or estimated and produce csDEG results as the output.

The methods described above are also reflections of the evolving methodology. csSAM, which was developed in 2010, is a pioneer in csDEG calling. It adopted a linear regression model and

Guanqun Meng is a PhD student in Epidemiology and Biostatistics at the Department of Population and Quantitative Health Sciences in Case Western Reserve University, School of Medicine. He is interested in *-omics* signal deconvolution methods in mixed high-throughput data.
Wen Tang is a biostatistician in the Department of Population and Quantitative Health Sciences at Case Western Reserve University, School of Medicine. She is interested in applied biostatistics and bioinformatics methods in team-science project research.
Emina Huang is the executive vice-chair of research and professor of surgery at the Department of Surgery in University of Texas Southwestern Medical Center. Her research interest is on inflammatory bowel disease and colorectal cancer genesis.
Ziyi Li is an assistant professor in the University of Texas MD Anderson Cancer Center. Her research focuses on developing statistical and machine learning methods and apply them to genomics, epigenetics and computational biology.
Hao Feng is an assistant professor in the Department of Population and Quantitative Health Sciences at Case Western Reserve University, School of Medicine. His main research interest is to develop statistical methods and computational tools for high-throughput bioinformatics data.
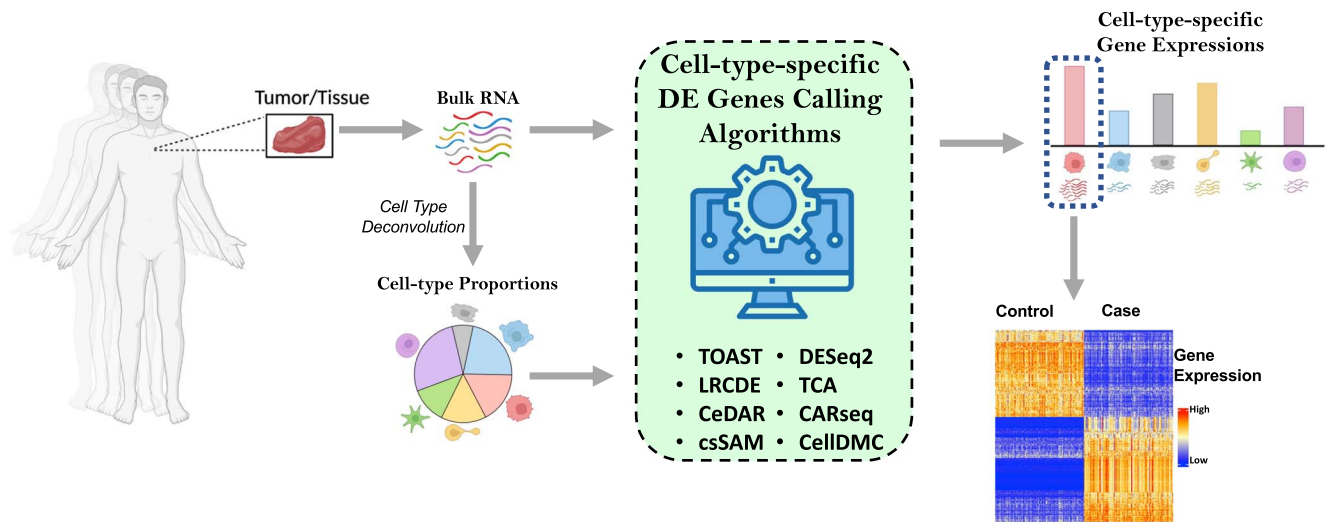
Figure 1. An overview of the csDEG detection experiment. Patient's samples are taken from biopsies and sequenced. Cell type compositions can be estimated by reference-free or reference-based deconvolution methods. With cell type proportions, bulk RNA-seq data, and samples' information as inputs, csDEG detection can be performed using several computational methods listed in the green box.

deconvoluted samples to test for csDEG using permutations. It was designed for gene expression microarray data initially. csSAM generated a limited set of distinct false discovery rate (FDR) values across the entire genome, due to the permutation design. It had low flexibility as it only focused on case–control comparisons and unable to take covariates. In 2014, DESeq2 was proposed. As one of the world-renowned DE detection methods, DESeq2 adopted a Negative Binomial distribution to model RNA-seq data, while allowing a generalized experimental design. Although DESeq2 was not designed for csDEG detection, it could achieve this goal by incorporating the cell type proportions as interaction covariates in the model. It is hereby included in our review as a representation of canonical DE algorithms that can also deliver csDEG results. LRCDE (2016) adopted a similar linear regression model to contrast expression coefficients from different phenotypes of interest. It represented another successful implementation of regression modeling. We see an upscaling in methodology development since 2018. CellDMC (2018) was designed for the cell type-specific analysis for DNA methylation data; nevertheless, its general framework is applicable to csDEG calling. In 2019, three algorithms, TCA, TOAST and HIRE were developed to deconvolute heterogeneous samples and identify csDEG. The modeling framework for TOAST was flexible, allowing users to incorporate additional covariates. TCA utilized tensor to deconvolve two-dimensional data into three-dimensional signals. HIRE targeted methylation data and applied a two-layer hierarchical model that captured the effects of the multiplicative cell composition on phenotype outcomes. With methylation data and covariates provided, HIRE directly output subject-level cell compositions, baseline methylation profiles and phenotype effects. Therefore, users do not need to provide cell type proportion information solved by external methods. In 2021, CARseq was proposed to use negative binomial distribution in csDEG analysis. It is a statistically rigorous approach because of a precise choice of distribution assumption and a sound modeling framework in RNA-seq count data. The trade-off is a long computing time caused by the iterative weighted least squares (IWLS) algorithm in parameter estimations. In 2022, CeDAR was developed to further incorporate csDEG correlation structures through a hierarchical tree. It demonstrated the advantage of using the information stored in

the tree structure to improve the csDEG calling process. As a reference for benchmark comparisons, Table 1 summarizes these methods in a chronological order.

We expect the csDEG detection, across the spectrum of research areas, will be widely adopted and continue to thrive. Therefore, a systematic review of currently available methods is in demand. Recently, efforts have been made to evaluate the sensitivities of some methods [29]. Despite being inclusive in its evaluation, the current research was conducted in a confined environment using only microarray real data, without simulation studies. Various aspects have not been thoroughly investigated, specifically for bulk RNA-seq data. Therefore, given the multifaceted aspects of csDEG calling and the existing wide adoption of bulk RNA-seq count data, a systematic and comprehensive review of the currently available methods will serve the immediate need of researchers who want to adopt such methods.

In this article, we systematically assessed all eight methods in RNA-seq experiments for csDEG detection performance. We designed a rigorous two-step simulation experiment to thoroughly assess their performances under various parameter choices. We benchmarked these methods under different scenarios, including various baseline expression levels, sample sizes, cell type compositions, expression level alterations and technical/biological dispersions. In real data analysis, we investigated their performances on three large bulk RNA-seq datasets on inflammatory bowel disease, lung cancer and autism studies. We summarized major factors affecting csDEG detection and inspected their performance stratified by expression levels. We paid particular attentions to methods' conservativeness in real data analyses and identified novel and classic pathways using consensus csDEG lists.

Data generative model and simulation

Given the multifaceted sources of randomness in cell type mixtures and sequencing steps, we designed a flexible and comprehensive simulation scheme as shown in Figure 2. The pipeline comprises three parts: (A) reference panel generation, (B) cell type proportion simulation and (C) weighted average and observed

Table 1. Summary of existing methods that are capable of identifying csDEG. Methods are ordered chronologically, with a brief summary of the algorithm, featured by pros and cons.

Method	Package/Year	Input	Algorithm	Features	
				Pros	Cons
Cell type-specific Significance Analysis of Microarray (csSAM)	csSAM/ 2010	Gene expression microarray data	Linear regression; deconvolute cases and controls separately. Inferences of csDEG are based on t-statistics of permutation.	The first model developed for csDEG calling.	1. Unable to adjust for batch effect/covariates. 2. Low flexibility, inflated false positive.
Differential gene expression based on NB[1] distribution (DESeq2)	DESeq2/ 2014	Gene expression RNA-seq data	Apply generalized NB[1] linear model and empirical Bayesian method to estimate the shrunk posterior of dispersion and LFC[2]. Adopt Wald tests under Normal distribution.	1. Stable parameter estimations for small replicates. 2. Widely applicable.	Not specifically for csDEG detection.
Linear Regression-based Cell type-specific Differential Expression (LRCDE)	lrcde/ 2016	General gene expression	Multivariate linear regressions: compare csDEG coefficients of different phenotypes. Inferences are based on two-sample t-test.	Highly sensitive when gene expression is low.	1. Underperforming in high expression genes. 2. Results significantly impacted by sample sizes and cell type proportions. Require moderate to large sample sizes.
Identification of Differentially Methylated Cell types (CellDMC)	EpiDISH/ 2018	DNA methylation	Multivariate linear regression solved by LSE.	Able to account for biological and technical effects.	1. Not specific for RNA-seq. 2. Unable to account for mean-variance relationship in count data.
Tools for the Analysis of heterogeneouS Tissues (TOAST)	TOAST/ 2019	Gene expression and methylation data	Linear model framework: incorporate cell type proportions, phenotype information, and subject-specific covariates.	1. Fast/flexible statistical testing. 2. Can accommodate different distributional assumptions, handle both count and continuous data.	
Tensor Composition Analysis (TCA)	TCA/ 2019	DNA methylation	Apply tensor to deconvolute 2D matrices into 3D tensors, which further allows statistical inference on variables of interest.	1. Good FDR controls. 2. High specificity/sensitivity.	1. Signals from abundant cell types are easier for detection than sparse cell types. 2. Cannot distinguish sources of signals if two cell types are associated.
Cell type-aware Analysis of RNA-seq (CARseq)	CARseq/ 2021	Gene expression RNA-seq data	NB regression with parameters estimated iteratively by IWLS. Inferences based on likelihood ratio test.	1. Variables divided into cell type-independent and cell type-specific categories for inference. 2. Handle low proportion cell types and small sample sizes.	1. Computational intensive. 2. Under-powered when csDEG occur in multiple abundant cell types.
CeDAR	TOAST/ 2022	Gene expression or methylation data	Stemmed from TOAST, further incorporating cell type DE/DM state correlations through hierarchical clustering.	Improved accuracies in abundant and associated cell types.	Not specific for RNA-seq data.

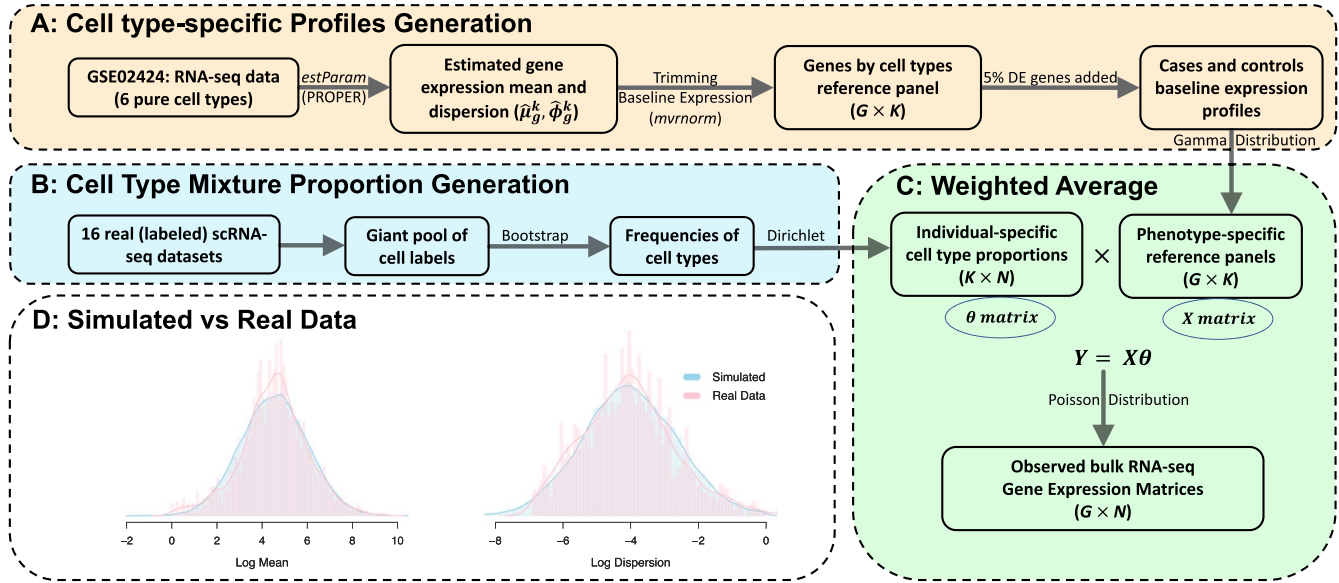[1] NB: negative binomial. [2] LFC: log fold change.

Figure 2. Bulk RNA-seq data simulation from mixtures of cell types. (**A**) Cell type profiles are generated from a multivariate log-normal distribution, with parameters estimated from pure cell line RNA-seq data. (**B**) Cell type proportions are generated from Dirichlet distribution, with parameters estimated from a collection of 16 well-labeled scRNA-seq datasets. (**C**) Gamma-Poisson compound is adopted to generate observed RNA-seq data, with dispersion embedded. (**D**) Simulated data mimic the real data well, reflected through the gene expression means and dispersion distributions.

expression simulation. These three steps are elucidated in the subsections below.

We use g to denote the gene index, with g from 1 to G. We use i as the sample index, with i from 1 to N. G and N represent the total number of genes and the total number of samples, respectively. Suppose the number of underlying cell types is K, indexed by k from 1 to K. Cell type-specific gene expression for each gene is denoted by $\mathbf{x}_g = (x_{g1}, ..., x_{gK})^T$. The sample-specific cell type mixing proportions is a vector denoted by $\boldsymbol{\theta}_i = (\theta_{1i}, \theta_{2i}..., \theta_{Ki})^T$. Naturally, it has the constraint $\sum_k \theta_{ki} = 1$, and $\theta_{ki} \geq 0$ for all k and i. For a specific gene g and a specific sample i, the observed bulk RNA-seq gene expression value, denoted by Y_{gi}, is a weighted sum of elements in $\mathbf{x}_g$, with weights given by $\boldsymbol{\theta}_i$.

Cell type expression profiles generation. We use a parametric model to generate pure cell type-specific gene expression profiles, with parameters estimated from real data. Here, we obtained a real RNA-seq dataset (GEO: GSE60424) [30], which includes six pure immune cell types (neutrophils, monocytes, B-cells, CD4 T cells, CD8 T cells and natural killer cells). Using the negative binomial modeling framework estimators provided by *PROPER* [31], we estimated the mean and dispersion for each gene at each cell type, for a total of $P = 1540$ genes. For each gene, denoted as p from 1 to P, the mean and dispersion estimations (in log scale) are $\hat{\boldsymbol{\mu}}_p = (\hat{\mu}_{p_1}, ..., \hat{\mu}_{p_K})^T$ and $\hat{\boldsymbol{\phi}}_p = (\hat{\phi}_{p_1}, ..., \hat{\phi}_{p_K})^T$. Real data have shown the associations of gene expression means and the dispersion of the same gene across different cell types. Therefore, properly accounting for such associations is not a trivial task. We adopt *multivariate normal distribution (MVN)* on $\hat{\boldsymbol{\mu}}_p$ and $\hat{\boldsymbol{\phi}}_p$, to account for the correlations among multiple cell types. In other words, we have $\hat{\boldsymbol{\mu}}_p \sim MVN(\boldsymbol{m}, \boldsymbol{\Sigma}_m)$ and $\hat{\boldsymbol{\phi}}_p \sim MVN(\boldsymbol{d}, \boldsymbol{\Sigma}_d)$. The *maximum likelihood estimation (MLE)* is then applied to get estimated $\hat{\boldsymbol{m}}, \hat{\boldsymbol{d}}, \hat{\boldsymbol{\Sigma}}_m$, and $\hat{\boldsymbol{\Sigma}}_d$. These parameters are adopted to simulate the underlying gene expression parameters for each simulated gene across all cell types:

$$\boldsymbol{\mu}_{g,K \times 1} \sim MVN(\hat{\boldsymbol{m}}, \hat{\boldsymbol{\Sigma}}_m)$$

$$\boldsymbol{\phi}_{g,K \times 1} \sim MVN(\hat{\boldsymbol{d}}, \hat{\boldsymbol{\Sigma}}_d),$$

where $\boldsymbol{\mu}_{g,K \times 1}$ and $\boldsymbol{\phi}_{g,K \times 1}$ are the simulated mean and dispersion parameters for one gene g. Then, the steps above are iterated for G times to obtain the mean (M) and dispersion (Φ) matrices for all genes and all cell types:

$$\mathbf{M}_{G \times K} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_G]^T; \quad \boldsymbol{\Phi}_{G \times K} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_G]^T.$$

In our simulations, we set $G = 30\,000$ and $K = 6$. Now we have two matrices of cell type-specific underlying gene expression parameters, $\mathbf{M}_{G \times K}$ and $\boldsymbol{\Phi}_{G \times K}$. A re-parameterized *Gamma Distribution* as shown below is to simulate the cell type-specific reference panel $\mathbf{X}_{G \times K}$.

$$\mathbf{X}_{G \times K} \sim \Gamma \{shape = \frac{1}{\exp(\boldsymbol{\Phi}_{G \times K})},$$

$$scale = \exp(\mathbf{M}_{G \times K}) \cdot \exp(\boldsymbol{\Phi}_{G \times K})\}$$

The biological dispersion is embedded in this *Gamma* step because it reflects the underlying variation of the group-level (case/control) reference panels. Additionally, the shape and scale parameters are transformed to reflect the mean and dispersion of its corresponding negative binomial distribution when it compounds with Poisson distribution (proof in Supplementary Section 1).

We assume cases and controls have different means but an identical dispersion. 5% of total genes are randomly selected to be differentially expressed, with half of the csDEG being upregulated and the remaining half being downregulated. The Log-Fold-Changes (LFCs) for csDEG are randomly drawn from $N(\mu_{LFC}, \sigma^2_{LFC})$ and $N(-\mu_{LFC}, \sigma^2_{LFC})$. We let μ_{LFC} and σ_{LFC} take values from 0.5 to 4 and 0.1 to 1, respectively, to benchmark these methods' performance under various scenarios.

Cell type mixture proportions generation. We leverage on well-labeled, annotated real single-cell RNA-seq (scRNA-seq) data to best mimic cell type proportions in simulation. Sixteen annotated real scRNA-seq datasets are obtained, and cell labels from these

studies are extracted (details in Supplementary Section 1.3). Cell types with very few cells are filtered out. We categorize cell labels and generate large pools of cell labels. Next, we can obtain the samples through bootstrap resampling. *Dirichlet Distribution* is adopted to estimate the *Dirichlet* parameter $\boldsymbol{\alpha}$ based on the bootstrapped samples. The cell type proportions are then simulated as follows:

$$\boldsymbol{\theta}_i \sim Dir(\boldsymbol{\alpha}).$$

In practice, we obtained two sets of parameters $\boldsymbol{\alpha}_C$ and $\boldsymbol{\alpha}_D$, respectively, for controls and cases to imitate the heterogeneity in proportions from different groups.

Observed RNA-seq count data generation. As outlined earlier, the observed expression level $\mathbf{Y}_i$ is a weighted average of pure cell type profiles. Therefore, we have:

$$\boldsymbol{w}_i = \mathbf{X}_{G \times K} \boldsymbol{\theta}_i$$

$$\mathbf{Y}_i | \boldsymbol{w}_i \sim Poisson(\boldsymbol{w}_i).$$

Here, both *Dirichlet* and *Poisson* steps reflect the random errors between individuals. Based on the mean expression $\boldsymbol{w}_i$, a Poisson distribution is adopted to simulate the observed RNA-seq count data. It also reflects the technical noise from the sequencing experiment, a different error source from biological dispersion introduced earlier by *Gamma* distribution.

In summary, this multi-step simulation design completes the compound structure equivalent to a negative binomial distribution (Supplementary Section 1). It has following advantages over using a negative binomial directly. First, the Gamma distribution captures the biological variations for the underlying true gene expression levels. Second, the Dirichlet distribution controls the variations on cell type proportions. Third, the weighted average of the underlying expression levels from Gamma distribution, evaluated by the weights generated from the Dirichlet step, naturally mimics the cell type mixture process. Fourth, the Poisson distribution captures the technical variations related to the randomness in the sequencing experiments. Overall, the benefit of this multi-step simulation design is to customize both the biological and technical noises, in addition to other parameters, for a better investigation on various aspects in csDEG identifications.

Overall snapshot

Given the estimated cell type proportions from real data, simulations are conducted to evaluate csDEG detection accuracy. We use the true discovery rate (TDR) as a main evaluation metric, which is defined as the percentage of true csDEG among significant top genes called by the algorithm. A high TDR is equivalent to a high precision. This reflects the practical consideration that researchers focus on the top identified genes; therefore, the accuracy among them matters the most. Canonical metrics, including receiver operating characteristic (ROC) curve, sensitivity and specificity, are also adopted.

We first conduct a baseline simulation study under a two-group comparison design, with a sample size *N*=100 for case and control groups, respectively. We set the proportion of DE genes at 5%, with Log Fold Change (LFC) mean and SD set to 1.5 and 0.2, respectively. There are 30 000 genes across six cell types simulated, repeated for a total of 20 times. Given the proportions solved by CIBERSORTx [32], they are treated as known values for a fair comparison and fed into all methods.

Figure 3A shows the averaged TDR results at one cell type and aggregated over all six cell types, respectively. Here, TOAST, CellDMC, TCA and CARseq have the highest TDR across different cell types. Especially, CellDMC and TOAST have almost identical performance. CeDAR is mostly comparable to TOAST and CARseq among the first several hundred top genes, but slightly reduced afterwards. The two traditional approaches, DESeq2 and csSAM show the lowest accuracy consistently. This is also within our expectation because they are either not designed to handle csDEG detection or adopt modeling for microarray data only. The sensitivity and type I error rate (also called False Positive rate, FPR, which equals 1-specificity), averaged across six cell types, are shown in Figure 3B. TCA provides the highest sensitivity, as well as a low sensitivity variability, across 20 simulations. It is followed by DEseq2 and CeDAR. In contrast, LRCDE shows a low sensitivity and inflated type I error rate. Figure 3C shows the relative locations of averaged sensitivity and FPR across simulations. The first-tier choice of method, favoring those with high sensitivity and low FPR, should contain the cluster of methods located near the upper left.

Similar conclusion can be drawn at additional simulations conducted at an exhaustive combinations of the number of genes at 10 000, 20 000, 30 000 and 40 000 and the number of cell types at 4, 5, 6, 7 and 8 (Supplementary Section 8.1 and 8.2).

Sample size and effect size of csDE

We next investigate the impacts of sample sizes and effect sizes on csDEG calling accuracy. The effect size is measured by the magnitude of LFC. Simulations are conducted under exhaustive combinations of sample sizes in 20, 50, 70, 100, and 200 per group and LFCs in 0.5, 1, 2, 3 and 4. These scenarios span common experimental conditions for sample sizes and effect sizes. TDR trends are shown in Figure 4 as a heatmap. All methods tend to perform well among very top-ranking genes identified (i.e. top 500). There is a decreasing trend in precision as we move down the top-ranking cutoffs. Under small effect size (i.e. LFC = 0.5), all methods tend to have poor precision even at large sample size. Larger sample sizes only improve TDR marginally. This suggests that the effect size plays a vital role in the csDEG detection problem: for those genes with moderate changes, deconvolution and detecting csDEG could be a challenging task itself. Increasing sample size can boost statistical power but its impact on precision is moderate. Meanwhile, we see a dramatic increase in TDR as the LFC increases. This improvement is evident, even at a small sample size. Overall, TOAST, CellDMC, TCA and CARseq exhibit favorable and stable performance in TDR.

Stratified expression

Previous study indicated a discrepancy in DE genes' detection at different gene expression ranges and library sizes [31]. Therefore, we stratify the genes by their expression values to investigate the methods' performances at each stratum. A total of nine gene expression strata are obtained: [0,10], [10,20], [20,40], [40, 80], [80,160], [160, 320), [320, 640), [640,1280), [1280, $+\infty$). Figure 5A shows the mean sensitivities across simulations of all eight methods at different strata, at various sample sizes and LFCs, and Figure 5B shows a stratum-specific boxplots sensitivity over 20 simulations. We generally observe poor sensitivity at low expression strata, especially for genes with expression values <20. When moving to higher strata, sensitivity increases and variability shrinks considerably. Almost all methods are sensitive to gene
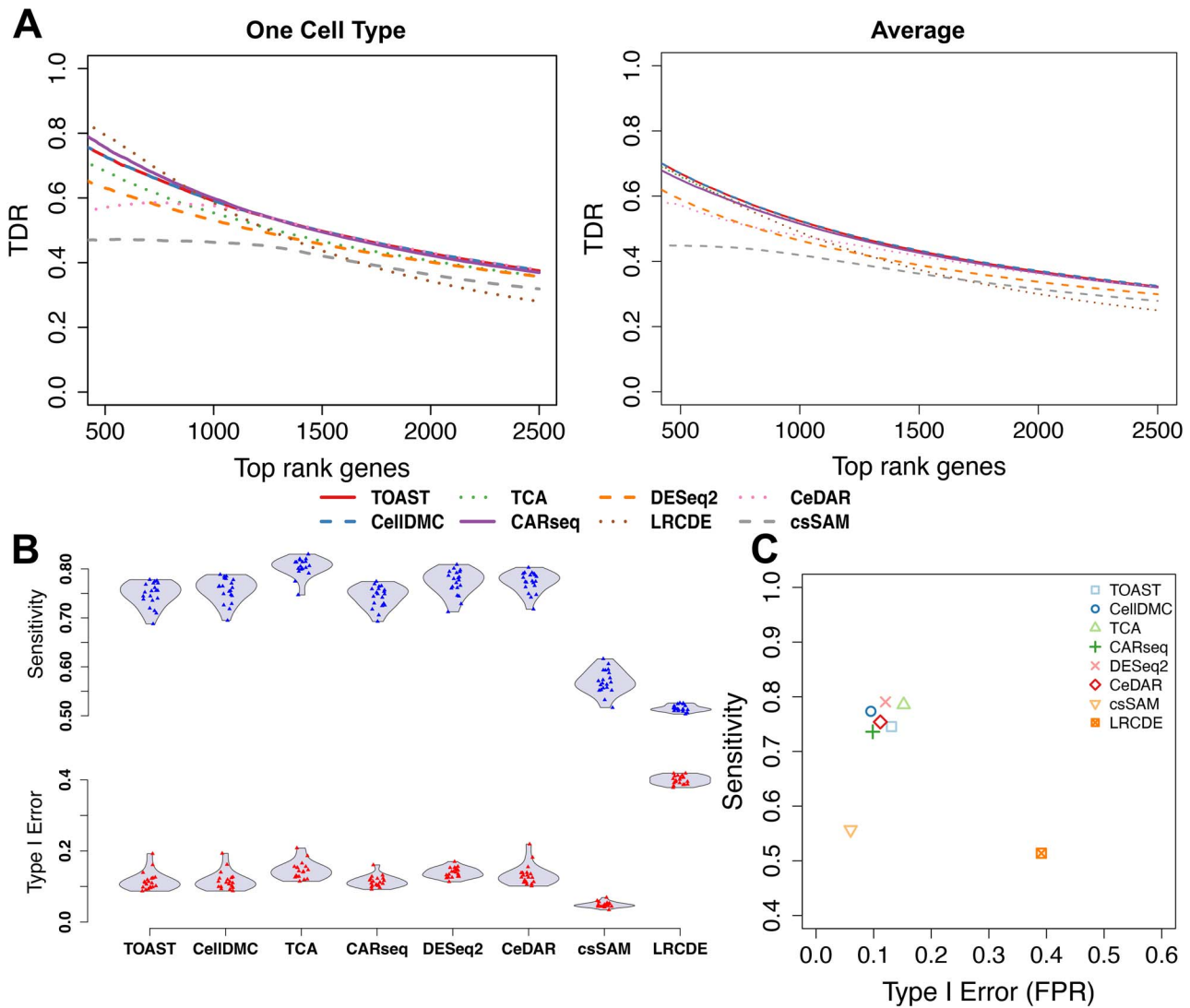
Figure 3. Comparisons of csDEG detection accuracy. (**A**) True discovery rate (TDR) is shown along the identified top-ranking genes in cell type six (left). The average TDR (right) is the aggregated results across six cell types. TDR is defined as the proportion of true csDEG, among all top-ranking genes, at a given cutoff. (**B**): Model-specific sensitivity and type I error rates averaged across six cell types. (**C**): Averaged type I error versus sensitivity for all eight models. $N=20$ simulations are conducted in each setting.

expression changes at high baseline expression levels. The benefit of sample size increase is not that profound among genes with low expression values (strata 1 and 2). This is because when the expression level is low (given the sequencing depth, only a handful of reads from a gene are sequenced), the technical noise shadows the real biological difference. In such situations, we are unlikely to have high probabilities of detecting csDEG while controlling for multiple testing. Two conclusions can be drawn here regarding sequencing depths and sample sizes: first, csDEG detection is more challenging than traditional DE analysis, and a larger sample size is preferred; second, a deep sequencing depth will contribute profoundly to genes on the boundary of discovery.

Cell type proportions

Our next benchmark investigates the impact of cell type proportions on csDEG detections. We select cell type proportions between 5%, 15%, 30% and 60%, to reflect the scenarios from a very minor cell type to a dominating one. Fixing the sample size per group at 100 and the LFC distribution mean to 1, we conduct 20 simulations and obtain the average performance. Figure 6 shows

the TDR as the cell type proportion increases, broken down by strata. As the proportion increase, we see a gradual improvement in performance across all methods. For those genes with high expression levels, having a marginal increase in proportion can contribute to the improvement considerably. On the other hand, for those genes with low expression levels (strata 1 and 2), the improvement in TDR is not as substantial as in other high-expression genes (strata 6 and above). This resonates with the previous results from stratified expression, showing the difficulties for genes with intrinsically low expression values. Besides the cell type proportion mean values, the power would decrease if the cell type proportions variations decrease, while holding the mean constant (Supplementary Section 7).

DE heterogeneity

We next investigate the impact of gene expression level alterations' heterogeneity on the csDEG detection accuracy. This set of simulations aim to assess if and how the variance of LFC's distribution across all genes would play a role. Similar to the scenarios above, we generate LFC and modify the baseline
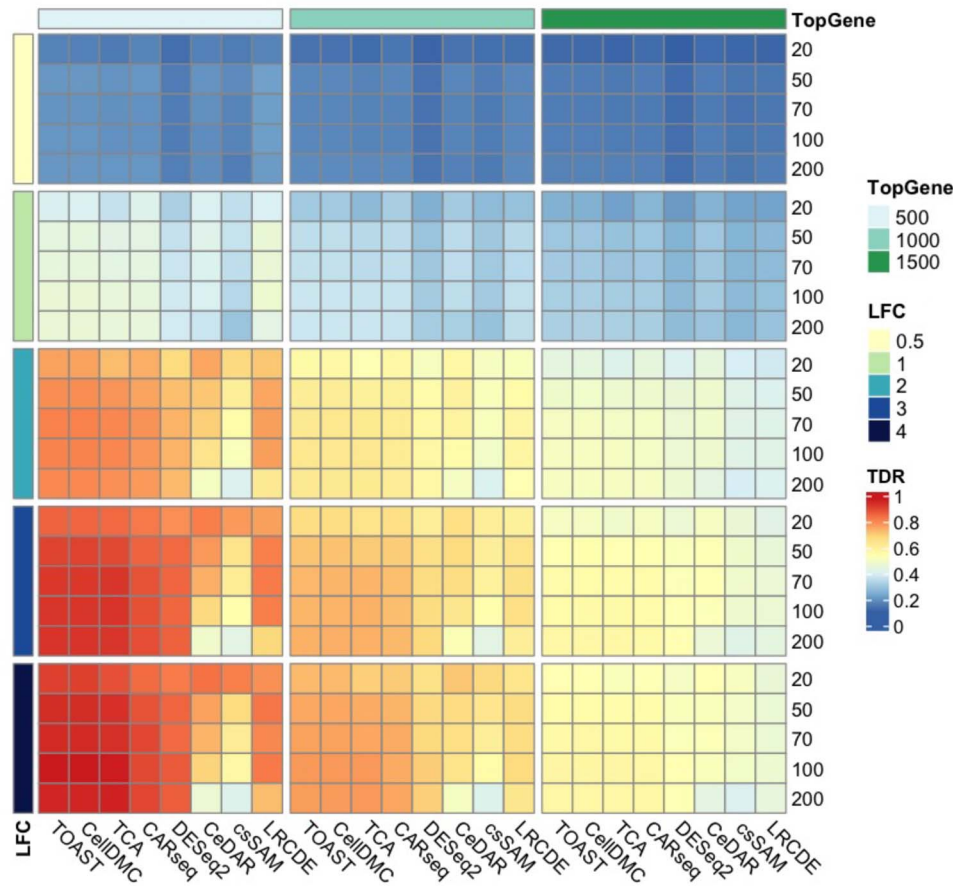
Figure 4. Heatmap showing TDR values at various combinations of sample sizes and LFCs combinations in simulation. The color bar on the left indicates the LFC, ranging from 0.5, 1, 2, 3 and 4. The numbers on the right side indicate the sample size per group, ranging from 20, 50, 70, 100 and 200. The color bar on the top indicates the cutoff for top-ranking genes at 500, 1000 and 1500.

expression level to create csDEG, but at various levels of the LFC's SD, ranging in 0.2, 0.4, 0.6, 0.8 and 1. A larger LFC's SD will lead to increased heterogeneity in simulated LFC distribution. The heatmap of TDR shows the simulation results in Supplementary Section 3.3. When the heterogeneity increases, we observe slightly reduced performance in TDR across all methods, although changes are minor to ignorable in general. Among them, LRCDE is most sensitive to DE heterogeneity. The remaining methods are relatively robust to various DE heterogeneity levels.

Running time

We benchmark the runtime for each method under the same computing environment. A total of 20 simulations are conducted, and the runtimes are averaged. As shown in Figure 7, TOAST has the fastest computational speed among all, due to a parallel back-end computing implementation and linear modeling structure. Most other methods can be completed within several minutes in a general personal computer environment. Generally, linear model-based methods such as TOAST, CellDMC and csSAM are fast. In contrast, CARseq would consume a significantly longer time due to its iterative algorithm.

Real data analysis

We apply csDEG calling methods on three large real datasets. The first dataset is the ileal transcriptome in pediatric inflammatory

bowel disease (IBD) (GEO: GSE57945) [33, 34]. This dataset includes a cohort of 359 treatment-naïve pediatric patients with Crohn's disease (CD, n = 213), ulcerative colitis (UC, n = 60) and healthy controls (n = 41). The ileal biopsies are obtained from colono-scopies diagnosis. This study focuses on the DE in UC patients and healthy controls younger than 17 years old. The second dataset is the bulk RNA-seq data in a large autism spectrum disorder (ASD) study [35, 36]. The study includes 251 samples of frontal/temporal cortex and cerebellum brain regions. Samples are from 48 ASD subjects versus 49 controls. Our analysis focuses on the cortex region, which contains 43 controls and 40 cases after filtering. The third dataset is the bulk RNA-seq data from The Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD). The outcomes are stages of pathologic cancer, which are further categorized into a binary variable, stage-I (n = 330) versus stage-II/III/IV (n = 271) status. In all analyses, we obtain the model-specific, as well as overlapping/consensus csDEG, and highlight the similarities and differences in pathways with extensive literature reviews. Maximum and minimal pairwise consensus csDEG for each cell type and across cell types, together with top relevant pathways, are shown in Supplementary Section 10.1–10.3, respectively, for UC, ASD and LUAD study. In the following, we first focus on the results from IBD study.

We apply the reference-free cell type deconvolution method, deconf [37], on the IBD dataset. Six different cell types are assumed for the deconvolution process. For each method, csDEG are identified based on sorted FDR values, and 5% of 28 052 total
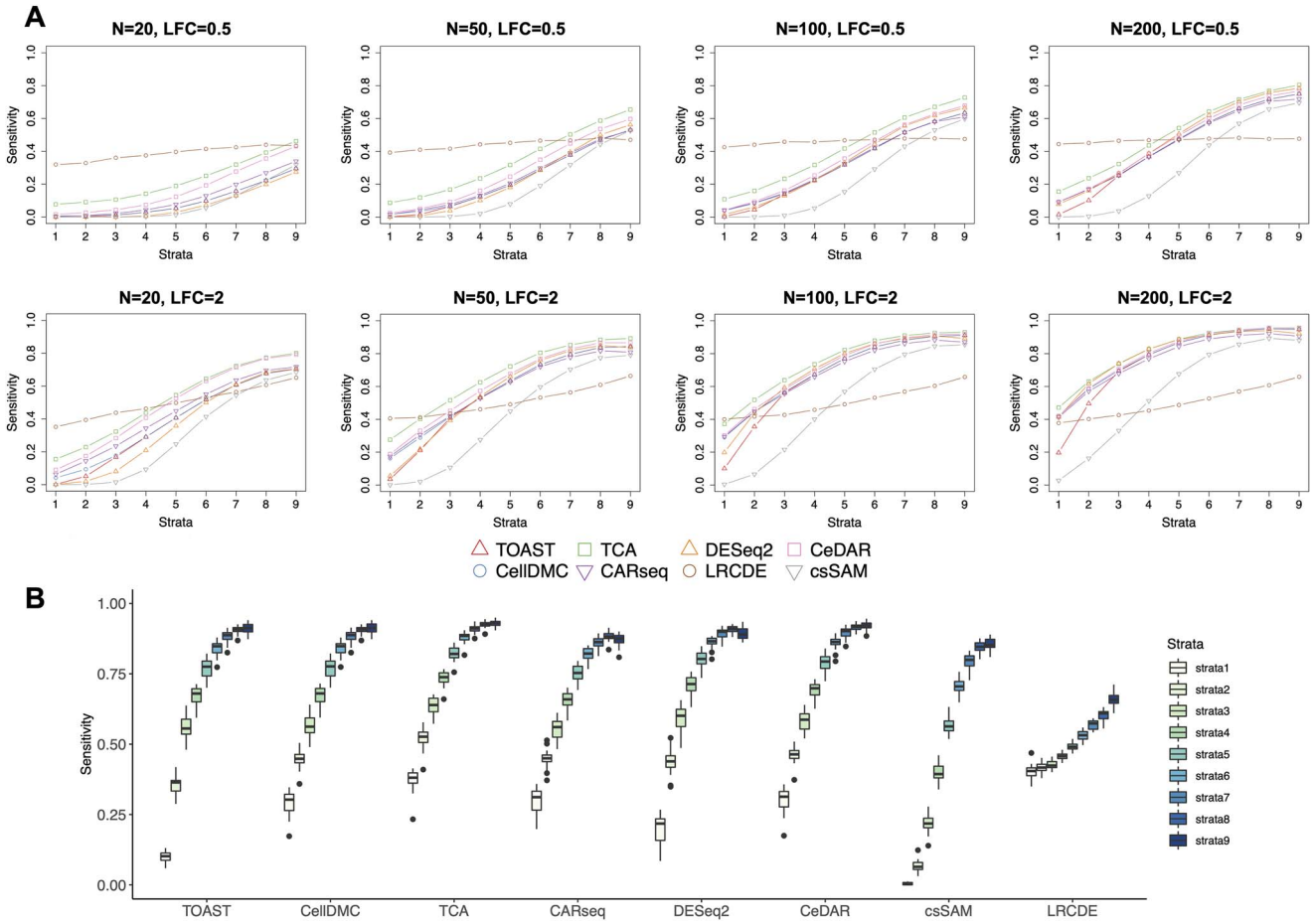
Figure 5. Sensitivity stratified by gene expression values. Nine strata are adopted: [0,10), [10,20), [20,40), [40, 80), [80,160), [160, 320), [320, 640), [640,1280), [1280, +∞) for strata 1-9, respectively. (**A**) Simulations at sample size N=20, 50, 100 and 200; and effect size LFC at 0.5 and 2. (**B**) Boxplots of sensitivities of 20 simulations, across cell types at each stratum, for sample size per group at N=100 and LFC following $N(2, 0.2)$.
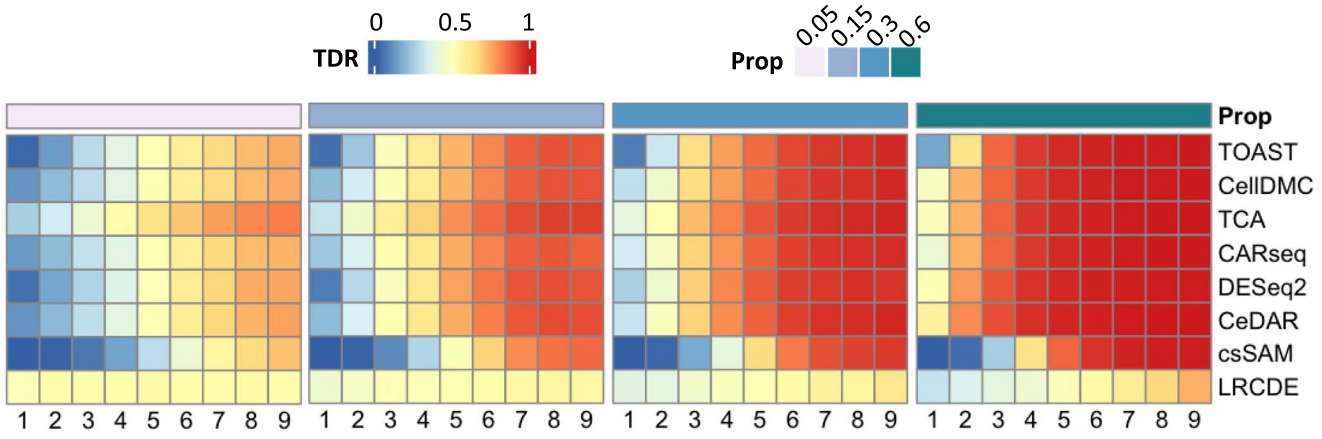

Figure 6. Heatmap of csDEG detection TDR values at various cell type proportions, with expression stratified. Cell type proportions studied: 5%, 15%, 30% and 60%, reflected through the top color bar of the color panels. Nine expression strata, labeled below each panel, have the same range as in Figure 5.

genes are identified as differentially expressed. The Venn diagram in Figure 8A shows the overlapping of identified csDEG across five methods. Although few genes are identified by all methods, a considerable amount of the identified genes are shared across at least two methods. For pairwise comparison, the maximal and minimal overlapping pairs are illustrated in Figure 8B and Figure 8C. Here, TOAST and LRCDE share the highest number of consensus csDEG (N = 671), whereas CARseq and DESeq2

have the lowest (N = 63). The consensus genes, respectively, for each scenario, are taken into Kyoto Encyclopedia of Genes and Genomes (KEGG) [38–40] pathway analysis, through Enrichr [41–43]. Top ranking IBD-relevant pathways are shown in the table of Figure 8D.

Although the exact pathology of IBD remains unclear, immune system impairment is one of the leading causes. Key pathways are identified from consensus csDEG identified from more than
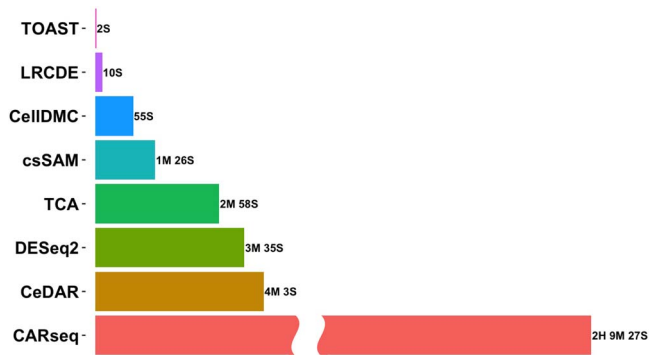
Figure 7. Runtime for all eight benchmarked methods under a general experimental scenario. A total of N=20 simulations are conducted, and runtimes are averaged. Computing environment: single node CPU with 64G RAM.

one methods. T-helper 17 (Th17), T-helper 1 (Th1) and T-helper 2 (Th2) are exhibited as the top-ranking pathways. These various CD4 T-cell subsets have been shown to play crucial roles in the pathogenesis of IBD, where Th1 has a role in Crohn's disease and Th2 has a role in UC [44]. Th17 is associated with IBD immunopathogenesisand autoinflammatory responses [45]. The current therapy is also directly targeted against Th1/Th17 cells differentiation in the intestinal mucosal inflammation, and the treatment has shown encouraging results [46–49]. In addition, chemokine signaling pathways are identified by both CARseq and DEseq2. Chemokines and their receptors inside gastrointestinal mucosal are important in regulating immune responses, mucosal homeostasis/inflammation, pathophysiological inflammation, physiological balance and colon cancer progressions [50,51]. CXCL8 chemokines, as well as CXCL1/CXCL2, as crucial pro-inflammatory factors, are elevated in UC pathogenesis [52]. Based on these recent findings, a novel IBD therapeutic approach is designed to target the chemokine family proteins [53].

We also identify classic pathways that have been adopted in UC treatment. For example, the JAK-STAT signaling pathway, which indicates the Janus kinase/signal transducer and activator (JAK-STAT) pathways, has been involved in many human disorders' pathogenesis. It is the target for a therapeutical agent, Tofacitinib, to treat severe UC [54,55]. The identified pathway of NAD^+, Nicotinamide adenine dinucleotide, is a critical coenzyme for redox reactions, which are central to energy metabolism and metabolic homeostasis; thus, it is essential for multiple cellular activities [56]. NAD^+ is critical for maintaining gut homeostasis in IBD patients [57,58]. Additional important pathways related to UC are covered in Supplementary Section 10.1.1.

Real data results also provide information for assessing whether a method is conservative or liberal. Model-specific csDEG levels are displayed in Figure 8F. Here, the model-specific csDEG level is defined as the ratio of the model-specific csDEG number over the number of consensus csDEG across multiple methods. Using this metric, a liberal method would have a higher value than the conservative one. DESeq2 yields more csDEG than others, whereas CellDMC and TOAST are relatively conservative. In general, for large-scale biomarker exploratory, liberal methods such as DESeq2, csSAM and TCA will produce more biomarkers for downstream analyses and wet-lab validations, though at the risk of inflated false positives. In contrast, conservative methods such as LRCDE, CellDMC and TOAST potentially have outstanding precision with a discount on sensitivity. Here for results related to significant genes overlapping and uniqueness, we used top 5%

as the cutoff to define significance. Similar conclusion, despite the minor changes in the order of methods, can be drawn by using a fixed FDR cutoff (Supplementary Section 10.1.8). Figure 8E shows the pairwise scatterplots of FDR values for common csDEG. The location of the scatter, relative to the diagonal line, can reflect the distribution of FDRs in comparison. CeDAR is relatively more aggressive than CellDMC, whereas TOAST and CellDMC have comparable performances due to the similar linear model framework.

The analysis for ASD dataset involves the cell type deconvolution based on a reference-based methods, ICeD-T [59]. Similarly, a binary outcome variable indicating controls and cases is created, and csDEG are identified (details in Supplementary Section 10.2). Our csDEG results support that neurodegenerations, neuronal cell loss and proinflammatory cytokines are strongly reflected in ASD patients [60,61]. Also, dysregulated neurotrophin signalings, brain-derived neurotrophic factor and their interactions with abnormal immune systems contribute to the development of ASD [62,63]. Several studies also suggest that a potent liver X receptor (LXR) agonist, TO901317, is a potential therapeutical strategy for ASD in relieving the difficulties in social interactions and restricted behaviors [64,65]. Our pathway analyses from consensus genes align with such findings.

For the third real data analysis on lung cancer, we adopt TIMER2.0 [66] to calculate cell proportions. This deconvolution algorithm requires standardized Transcripts Per Million inputs, and six immune cell types are B cell, T cell CD4, T cell CD8, neutrophil, macrophage and myeloid dendritic. From our pathway analyses results, we are able to provide supporting evidence to the previous discovery that extracellular RNAs (exRNAs) could facilitate NETs and induce lung cancer oncogenesis [67]. Supplementary Section 10.3 contains additional results from nutritional and dietary perspective interpretations, for consensus csDEG identified.

In both UC and ASD datasets analyses, TOAST and LRCDE share the highest number of consensus csDEGs, and CARseq and DESeq2 share the lowest number of consensus csDEGs. In UC and LUAD analysis settings, we observe that CellDMC and TOAST are more conservative than CeDAR. Across all three datasets, we also notice that csSAM and DESeq2 identify more model-specific csDEGs than other models. They are likely to produce inflated FPRs, when combining the observation with simulation findings. On the contrary, TOAST and CellDMC are more conservative. TCA, CeDAR and CARseq are at moderate levels of uncovering model-specific csDEG.

Discussion

We systematically evaluate several novel and classic methods to identify csDEG. As the data analysis granularity of DE analysis shifts from bulk to cell type level, these methods are gaining popularity and being widely adopted. Our investigation of their performances, under various simulation settings, provides researchers tangible guidance for method selection.

We design a rigorous two-step data generation framework in simulation, which helps us control the level of various parameters, including sample sizes, effect sizes, effect size heterogeneity, cell type numbers, cell type abundances and csDEG abundances. In the context of stratified gene expression, we investigate all methods and report their performances. Primary metrics include TDR, ROC, sensitivity, specificity, FPR and runtime, with particular emphasis on TDR. This underscores the utility of top-ranking biomarkers in the clinical research setting.
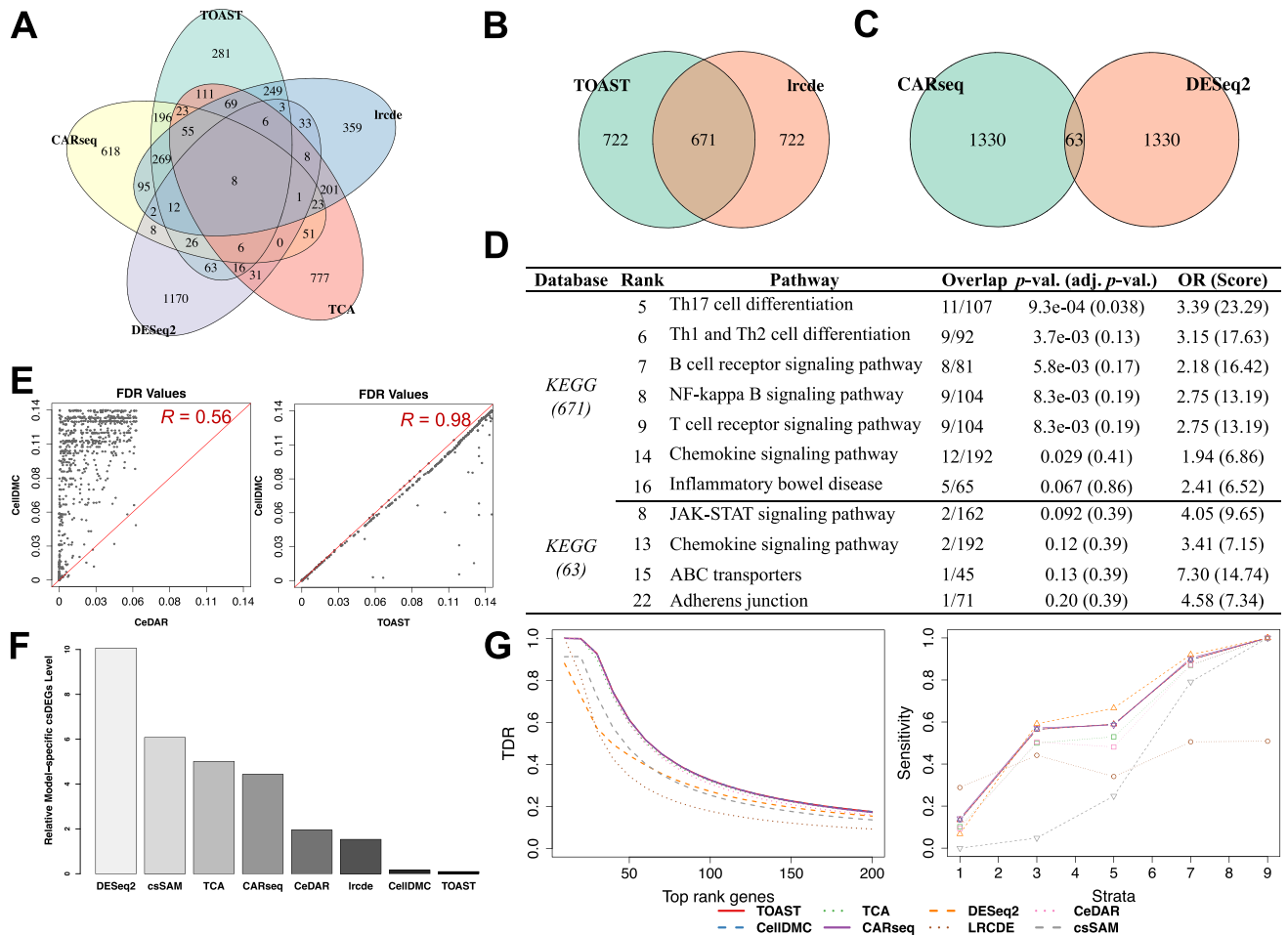
Figure 8. Real data analysis for a set of IBD RNA-seq dataset (GEO: GSE57945). (**A**) Venn diagram showing the overlapped csDEG from CARseq, DESeq2, TOAST, TCA and LRCDE. (**B**) and (**C**) illustrate the maximal and minimal pairwise overlapping csDEG among all the methods. (**D**) Relevant pathways identified by the significant and overlapped csDEG in (**B**) and (**C**). (**E**) Scatterplots of computed FDR values across two pairwise comparisons. Correlation coefficients are shown in red color. (**F**) For each method, the model-specific csDEG level is shown. The higher value indicates a more liberal model. The model-specific csDEG level is defined as the ratio between model-specific csDE genes number and consensus-called genes number. (**G**) shows model accuracies based on pseudo-bulk RNA-seq data derived from a real single-cell dataset. The left panel shows the TDR of the top-ranking genes and the right panel shows sensitivities across strata.

Based on the baseline simulation and TDR metric, we notice that across all six cell types, TOAST, CellDMC and CARseq provide the highest accuracy. CellDMC and TOAST have almost identical performance in terms of TDR (Figure 3) and FDR correlations (Figure 8), because both adopt a linear regression framework. Although CARseq has slightly lower TDR, its statistical model is rigorous and reflects the most sophisticated modeling in count data. Classic methods like DESeq2 and csSAM, although can be customized to solve this problem, are not as good as those recently developed ones. LRCDE shows unstable performances across the top selected genes partly due to its liberal behavior (Supplementary Figure S3, LRCDE is removed due to its over-sized scale).

From the sample size and effect size perspectives, a small effect size plays a strong and detrimental effect on csDEG calling. We conclude that the csDEG calling is more challenging than the classic DE problem in bulk data, even with a newly developed models. There is barely any model that has an ideal performance (TDR < 30%) under a small effect size (LFC = 0.5). At a reasonably large effect size, CARseq, TOAST, CellDMC and TCA are in the tier-1 group in testing accuracy (reflected by TDR) and should be

considered favorably by researchers. Those top-ranking identified genes have higher chances of being true positives.

The gene expression level is an underlying confounding factor for the model performance, so we have stratified it in our study. For most methods, there is a clear increasing trend in csDEG calling sensitivity as we move from a low stratum (low expression) to a high stratum (high expression). The intuition is that detecting csDEG is a relatively easy task for highly expressed genes. LRCDE is an aggressive model that tends to generate extremely small FDR values, which lead to elevated false-positive rates despite boosted sensitivities. It is also the reason for its acceptable sensitivities even at low expression strata (<80) and small LFC. Sensitivity is also more susceptible to changes in sample size at higher expression strata for all models except LRCDE. For TOAST, CellDMC, CARseq, TCA and DESeq2, the sensitivity is more reactive to effect sizes at lower expression strata and large sample sizes. Therefore, when both sample sizes and expression values are low, we typically have insufficient power to detect true csDEG. In this case, technical sequencing noises dominate true signals, making it a challenging task to accomplish. Thus, we recommend deeper sequencing depths to rescue those low expression genes and more

attentions to these identified biomarkers with high expression values.

Through the investigations of cell type proportions, we found that cell type proportions are positively associated with models' performances. This association is substantial, especially for genes with moderate to high expression values. Genes in low strata are resilient to such change, likely due to the dominance of noise. For all models, especially TOAST, CellDMC, TCA, CARseq and CeDAR, TDR can be considerably improved by cell type abundance increase at a high gene expression stratum, than the same proportion increase at a low gene expression stratum. In contrast, LRCDE has a subpar performance, even at increased cell type proportion settings.

It is worth noting that the csDEG calling is a challenging task, whose accuracy can be hampered by several key factors. First, small effect size at cell type resolution would negatively impact testing accuracy. For example, at small effect size (LFC = 0.5), models would all suffer from reduced precision (~20% TDR at 500 top discovered genes). It is likely that the technical noise would overwhelm the biological variation and lead to less favorable accuracy. Second, available models for csDEG detection are most suitable for moderate to high expression genes. Genes in the low strata, even if they have high effect size (LFC = 2), would suffer from low sensitivity issues. Researchers will encounter insufficient power for low expression genes; therefore, a high sequencing depth is recommended. Additionally, a higher sequencing depth can partly alleviate the small effect size problem described earlier. For example, the sensitivity would reach ~60% for most models at a high strata (e.x. strata 6–9). Third, current models do not provide the optimal solution to handle repeatedly measured samples within each subject. For those longitudinally observed data with correlated reference panels, the csDEG detection is a research problem still unsolved.

Key Points

- Cell type-specific differentially expressed genes (csDEG) analysis is successful at dissecting bulk RNA-seq data and identifying biomarkers in a finer resolution.
- Effect size, baseline expression level and cell type composition are the leading factors affecting csDEG calling accuracy.
- CARseq, TOAST, CellDMC and TCA are the most reliable methods in terms of precision and sensitivity.
- Insufficient power can be expected for low expression genes. Larger sample size is needed compared with traditional DE analysis.
- csDEG is a challenging task itself, with room to improve to properly handle low signal-to-noise ratio and low expression genes.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Author contributions statement

Z.L. and H.F. conceived the experiments, G.M. conducted the simulations and real data analysis, W.T. compiled results and the figures, E.H. identified real datasets and guide analysis interpretations, H.F. and G.M. wrote the manuscript. All authors reviewed the manuscript.

Availability of data/method

- Simulation Dataset: Next generation sequencing of human immune cells (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60424)
- Parameter Estimation: PROPER package (https://bioconductor.org/packages/release/bioc/html/PROPER.html)
- TOAST/CeDAR (https://bioconductor.org/packages/release/bioc/html/TOAST.html)
- CellDMC: EpiDISH package (https://bioconductor.org/packages/EpiDISH/)
- TCA (https://cran.r-project.org/web/packages/TCA/vignettes/tca-vignette.html)
- CARseq (https://github.com/Sun-lab/CARseq)
- DESeq2 (https://bioconductor.org/packages/release/bioc/html/DESeq2.html)
- csSAM (https://github.com/shenorrLab/csSAM)
- LRCDE (https://github.com/ERGlass/lrcde.dev)
- deconf (http://web.cbio.uct.ac.za/~renaud/CRAN/#)
- CIBERSORTx (https://github.com/IOBR/IOBR)
- TIMER2.0 (https://github.com/omnideconv/immunedeconv)
- Inflammatory Bowel Disease Real Data: Core Ileal Transcriptome in Pediatric Crohn Disease (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57945)
- ASD Real Data: https://www.synapse.org/#!Synapse:syn8365527
- TCGA-LUAD: https://jhubiostatistics.shinyapps.io/recount/

References

1. Lin M, Pedrosa E, Shah A, *et al*. Rna-seq of human neurons derived from ips cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders. *PLoS one* 2011;**6**(9):e23356.
2. Leidinger P, Backes C, Deutscher S, *et al*. A blood based 12-mirna signature of Alzheimer disease patients. *Genome Biol* 2013;**14**(7):1–16.
3. Lau P, Bossers K, Janky R, *et al*. Alteration of the micro RNA network during the progression of Alzheimer's disease. *EMBO Mol Med* 2013;**5**(10):1613–34.
4. Magistri M, Velmeshev D, Makhmutova M, *et al*. Transcriptomics profiling of Alzheimer's disease reveal neurovascular defects,

altered amyloid-β homeostasis, and deregulated expression of long noncoding rnas. *J Alzheimers Dis* 2015;**48**(3):647–65.

5. Kang Y, Zhou Y, Li Y, *et al*. A human forebrain organoid model of fragile x syndrome exhibits altered neurogenesis and highlights new treatment strategies. *Nat Neurosci* 2021;**24**(10):1377–91.

6. Jiao B, Wang M, Feng H, *et al*. Downregulation of top2 modulates neurodegeneration caused by ggggcc expanded repeats. *Hum Mol Genet* 2021;**30**(10):893–901.

7. Rahman MR, Petralia MC, Ciurleo R, *et al*. Comprehensive analysis of rna-seq gene expression profiling of brain transcriptomes reveals novel genes, regulators, and pathways in autism spectrum disorder. *Brain Sci* 2020;**10**(10):747.

8. Zhang F, Hammack C, Ogden SC, *et al*. Molecular signatures associated with zikv exposure in human cortical neural progenitors. *Nucleic Acids Res* 2016a;**44**(18):8610–20.

9. Wilson JA, Prow NA, Schroder WA, *et al*. Rna-seq analysis of chikungunya virus infection and identification of granzyme a as a major promoter of arthritic inflammation. *PLoS Pathog* 2017;**13**(2):e1006155.

10. Wang Y, Lupiani B, Reddy S, *et al*. Rna-seq analysis revealed novel genes and signaling pathway associated with disease resistance to avian influenza virus infection in chickens. *Poult Sci* 2014;**93**(2):485–93.

11. Hwang Y, Kim J, Shin J, *et al*. Gene expression profiling by mrna sequencing reveals increased expression of immune/inflammation-related genes in the hippocampus of individuals with schizophrenia. *Transl Psychiatry* 2013;**3**(10):e321–1.

12. Sarvestani SK, Signs S, Hu B, *et al*. Induced organoids derived from patients with ulcerative colitis recapitulate colitic reactivity. *Nat Commun* 2021;**12**(1):1–18.

13. Baumgart M, Priebe S, Groth M, *et al*. Longitudinal rna-seq analysis of vertebrate aging identifies mitochondrial complex i as a small-molecule-sensitive modifier of lifespan. *Cell Syst* 2016;**2**(2):122–32.

14. Wooff Y, Cioanca AV, Chu-Tan JA, *et al*. Small-medium extracellular vesicles and their mirna cargo in retinal health and degeneration: mediators of homeostasis, and vehicles for targeted gene therapy. *Front Cell Neurosci* 2020;**14**:160.

15. Beane J, Vick J, Schembri F, *et al*. Characterizing the impact of smoking and lung cancer on the airway transcriptome using rna-seq. *Cancer Prev Res* 2011;**4**(6):803–17.

16. Li Y, Xiao X, Ji X, *et al*. Rna-seq analysis of lung adenocarcinomas reveals different gene expression profiles between smoking and nonsmoking patients. *Tumor Biol* 2015;**36**(11):8993–9003.

17. Zheng GX, Terry JM, Belgrader P, *et al*. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;**8**(1):1–12.

18. Yu X, Chen Y, Conejo-Garcia JR, *et al*. Estimation of immune cell content in tumor using single-cell rna-seq reference data. *BMC Cancer* 2019;**19**(1):1–11.

19. Deng W, Ma Y, Su Z, *et al*. Single-cell rna-sequencing analyses identify heterogeneity of cd8+ t cell subpopulations and novel therapy targets in melanoma. *Mol Ther Oncolytics* 2021;**20**:105–18.

20. Jin C, Chen M, Lin D-Y, *et al*. Cell-type-aware analysis of rna-seq data. *Nat Comput Sci* 2021;**1**(4):253–61.

21. Li Z, Wu Z, Jin P, *et al*. Dissecting differential signals in high-throughput data from complex tissues. *Bioinformatics* 2019b;**35**(20):3898–905.

22. Chen L, Li Z, Wu H. Cedar: incorporating cell type hierarchy improves cell type specific differential analyses in bulk omics data. *bioRxiv* 2022. https://doi.org/10.1101/2022.07.09.499410.

23. Zheng SC, Breeze CE, Beck S, *et al*. Identification of differentially methylated cell types in epigenome-wide association studies. *Nat Methods* 2018;**15**(12):1059–66.

24. Glass ER, Dozmorov MG. Improving sensitivity of linear regression-based cell type-specific differential expression deconvolution with per-gene vs. global significance threshold. *BMC Bioinformatics* 2016;**17**:163–75.

25. Rahmani E, Schweiger R, Rhead B, *et al*. Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nat Commun* 2019;**10**(1):1–11.

26. Shen-Orr SS, Tibshirani R, Khatri P, *et al*. Cell type–specific gene expression differences in complex tissues. *Nat Methods* 2010;**7**(4):287–9.

27. Luo X, Yang C, Wei Y. Detection of cell-type-specific risk-cpg sites in epigenome-wide association studies. *Nat Commun* 2019;**10**(1):1–12.

28. Love M, Anders S, Huber W. Differential analysis of count data–the deseq2 package. *Genome Biol* 2014;**15**(550):10–1186.

29. Jaakkola MK, Elo LL. Estimating cell type-specific differential expression using deconvolution. *Brief Bioinform* 2022;**23**(1):bbab433.

30. Linsley PS, Speake C, Whalen E, *et al*. Copy number loss of the interferon gene cluster in melanomas is linked to reduced t cell infiltrate and poor patient prognosis. *PloS one* 2014;**9**(10):e109760.

31. Wu H, Wang C, Wu Z. Proper: comprehensive power evaluation for differential expression using rna-seq. *Bioinformatics* 2015;**31**(2):233–41.

32. Newman AM, Steen CB, Liu CL, *et al*. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* 2019;**37**(7):773–82.

33. Haberman Y, Tickle TL, Dexheimer PJ, *et al*. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. *J Clin Invest* 2014;**124**(8):3617–33.

34. Loberman-Nachum N, Sosnovski K, Di Segni A, *et al*. Defining the celiac disease transcriptome using clinical pathology specimens reveals biologic pathways and supports diagnosis. *Sci Rep* 2019;**9**(1):16163.

35. Gandal MJ, Zhang P, Hadjimichael E, *et al*. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* 2018;**362**(6420):eaat8127.

36. Parikshak NN, Swarup V, Belgard TG, *et al*. Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* 2016;**540**(7633):423–7.

37. Repsilber D, Kern S, Telaar A, *et al*. Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics* 2010;**11**(1):27.

38. Kanehisa M, Goto S. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**(1):27–30.

39. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci* 2019;**28**(11):1947–51.

40. Kanehisa M, Furumichi M, Sato Y, *et al*. Kegg: integrating viruses and cellular organisms. *Nucleic Acids Res* 2020;**49**(D1):D545–51.

41. Chen EY, Tan CM, Kou Y, *et al*. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics* 2013;**14**:128.

42. Kuleshov MV, Jones MR, Rouillard AD, *et al*. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;**44**(W1):W90–7.

43. Xie Z, Bailey A, Kuleshov MV, *et al*. Gene set knowledge discovery with enrichr. *Current Protocols* 2021;**1**(3):e90.

44. Zenewicz LA, Antov A, Flavell RA. Cd4 t-cell differentiation and inflammatory bowel disease. *Trends Mol Med* 2009;**15**(5):199–207.

45. Zhao J, Lu Q, Liu Y, *et al.* Th17 cells in inflammatory bowel disease: cytokines, plasticity, and therapies. *J Immunol Res* 2021;**2021**:8816041.

46. Liu Z-J, Yadav P-K, Su J-L, *et al.* Potential role of th17 cells in the pathogenesis of inflammatory bowel disease. *World J Gastroenterol* **15**:5784–8.

47. Su J, Chen T, Ji X-Y, *et al.* IL-25 downregulates Th1/Th17 immune response in an IL-10-dependent manner in inflammatory bowel disease. *Inflamm Bowel Dis* 2013;**19**(4):720–8.

48. Zhu Q, Zheng P, Zhou J, *et al.* Andrographolide affects th1/th2/th17 responses of peripheral blood mononuclear cells from ulcerative colitis patients. *Mol Med Rep* 2018;**18**(1): 622–6.

49. Arctigenin exerts anti-colitis efficacy through inhibiting the differentiation of th1 and th17 cells via an mtorc1-dependent pathway. *Biochem Pharmacol* 2015;**96**:323–36.

50. Zimmerman NP, Vongsa RA, Wendt MK, *et al.* Chemokines and chemokine receptors in mucosal homeostasis at the intestinal epithelial barrier in inflammatory bowel disease. *Inflamm Bowel Dis* 2008;**14**(7):1000–11.

51. Singh UP, Singh NP, Murphy EA, *et al.* Chemokine and cytokine levels in inflammatory bowel disease patients. *Cytokine* 2016;**77**:44–9.

52. Zhu Y, Yang S, Zhao N, *et al.* Cxcl8 chemokine in ulcerative colitis. *Biomed Pharmacother* 2021;**138**:111427.

53. Camba-Gómez M, Arosa L, Gualillo O, *et al.* Chemokines and chemokine receptors in inflammatory bowel disease: recent findings and future perspectives. *Drug Discov Today* 2022; **27**(4):1167–75.

54. Coskun M, Salem M, Pedersen J, *et al.* Involvement of jak/stat signaling in the pathogenesis of inflammatory bowel disease. *Pharmacol Res* 2013;**76**:1–8.

55. Sandborn WJ, Ghosh S, Panes J, *et al.* Tofacitinib, an oral Janus kinase inhibitor, in active ulcerative colitis. *N Engl J Med* 2012;**367**(7):616–24.

56. Covarrubias AJ, Perrone R, Grozio A, *et al.* Nad+ metabolism and its roles in cellular processes during ageing. *Nat Rev Mol Cell Biol* 2021;**22**(2):119–41.

57. Navarro MN, Gómez de las Heras MM, Mittelbrunn M. Nicotinamide adenine dinucleotide metabolism in the immune response, autoimmunity and inflammageing. *Br J Pharmacol* 2022;**179**(9):1839–56.

58. Han X, Uchiyama T, Sappington PL, *et al.* Nad+ ameliorates inflammation-induced epithelial barrier dysfunction in cultured enterocytes and mouse ileal mucosa. *J Pharmacol Exp Ther* 2003;**307**(2):443–9.

59. Wilson DR, Jin C, Ibrahim JG, *et al.* Iced-t provides accurate estimates of immune cell abundance in tumor samples by allowing for aberrant gene expression patterns. *J Am Stat Assoc* 2020;**115**(531):1055–65.

60. Kern JK, Geier DA, Sykes LK, *et al.* Evidence of neurodegeneration in autism spectrum disorder. *Transl Neurodegeneration* 2013;**2**(1):1–6.

61. Jęśko H, Cieślik M, Gromadzka G, *et al.* Dysfunctional proteins in neuropsychiatric disorders: from neurodegeneration to autism spectrum disorders. *Neurochem Int* 2020;**141**:104853.

62. Ohja K, Gozal E, Fahnestock M, *et al.* Neuroimmunologic and neurotrophic interactions in autism spectrum disorders: relationship to neuroinflammation. *Neuromolecular Med* 2018; **20**(2):161–73.

63. Hicks SD, Middleton FA. A comparative review of microrna expression patterns in autism spectrum disorder. *Front Psych* 2016;**7**:176.

64. Cai Y, Zhong H, Li X, *et al.* The liver x receptor agonist to901317 ameliorates behavioral deficits in two mouse models of autism. *Front Cell Neurosci* 2019;**13**:213.

65. Zhang J-X, Zhang J, Li Y. Liver x receptor-β improves autism symptoms via downregulation of β-amyloid expression in cortical neurons. *Ital J Pediatr* 2016b;**42**(1):1–9.

66. Sturm G, Finotello F, Petitprez F, *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* 2019;**35**(14):i436–45.

67. Li Y, Yang Y, Gan T, *et al.* Extracellular rnas from lung cancer cells activate epithelial cells and induce neutrophil extracellular traps. *Int J Oncol* 2019a;**55**(1):69–80.

68. Karosi T, Montinaro V, Dann SM, *et al.* Role of th17 cells in the pathogenesis of human IBD. *ISRN Inflamm* 2014;928461.

69. Nanjappa MS, Voyiaziakis E, Pradhan B, *et al.* Use of selective serotonin and norepinephrine reuptake inhibitors (snris) in the treatment of autism spectrum disorder (ASD), comorbid psychiatric disorders and ASD-associated symptoms: a clinical review. *CNS Spectr* 2022;**27**(3):290–7.