PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data

Dongyuan Song [1] and Jingyi Jessica Li [2,3,4,*]

Abstract

In the investigation of molecular mechanisms underlying cell state changes, a crucial analysis is to identify differentially expressed (DE) genes along a continuous cell trajectory, which can be estimated by pseudotime inference from single-cell RNA-sequencing (scRNA-seq) data. However, existing methods that identify DE genes based on inferred pseudotime do not account for the uncertainty in pseudotime inference. Also, they either have ill-posed p-values that hinder the control of false discovery rate (FDR) or have restrictive models that reduce the power of DE gene identification. To overcome these drawbacks, we propose PseudotimeDE, a robust method that accounts for the uncertainty in pseudotime inference and thus identifies DE genes along cell pseudotime with well-calibrated p-values. PseudotimeDE is flexible in allowing users to specify the pseudotime inference method and to choose the appropriate model for scRNA-seq data. Comprehensive simulations and real-data applications verify that PseudotimeDE provides well-calibrated p-values essential for controlling FDR and downstream analysis and that PseudotimeDE is more powerful than existing methods to identify DE genes.

Introduction

In recent years, single-cell RNA-sequencing (scRNA-seq) technologies have undergone rapid development to dissect transcriptomic heterogeneity and to discover cell types or states in complex tissues [1, 2]. Embracing the capacity to measure transcriptomes of numerous cells simultaneously, scRNA-seq provides a powerful means to capture continuous cell-state transition across cells, and it has been used to study key cellular processes such as immune response [3] and cell development [4]. For example, a study of human fibroblasts identified distinct fibroblast subtypes responsible for mediating inflammation or tissue damage in arthritis [5]; a study of maternal-fetal interface tissue revealed new cell states and the importance of this tissue in maternal immune

[1] Bioinformatics Interdepartmental Ph.D. Program, University of California, Los Angeles, CA 90095-7246
[2] Department of Statistics, University of California, Los Angeles, CA 90095-1554
[3] Department of Human Genetics, University of California, Los Angeles, CA 90095-7088
[4] Department of Computational Medicine, University of California, Los Angeles, CA 90095-1766
[*] To whom correspondence should be addressed. Email: jli@stat.ucla.edu

tolerance of paternal antigens [6]; a study of thymic development elucidated new principles of naïve T cell repertoire formation [7].

Pseudotime inference, also known as trajectory inference, is one of the most thriving scRNA-seq data analysis topics. The concept of "pseudotime" was first proposed in 2014 [8], and since then, more than 40 pseudotime inference methods have been developed [9]. Pseudotime inference aims to infer the ordering of cells along a lineage based on the cells' gene expression profiles measured by scRNA-seq, and the inferential target is "pseudotime," a time-like variable indicating the relative position a cell takes in a lineage. By establishing a temporal dimension in a static scRNA-seq dataset, pseudotime inference allows the probing of individual genes' expression dynamics along with continuous cell-state changes. If a gene's mean expression changes along pseudotime, the gene is referred to as differentially expressed (DE) and is likely to play an important role in the underlying cellular process that gives rise to the pseudotime. Identifying DE genes is the most crucial analysis after pseudotime inference because genes are the most fundamental functional units for understanding biological mechanisms.

Several methods have been developed to identify DE genes along inferred cell pseudotime. Popular pseudotime inference methods—TSCAN [10], Slingshot [11], Monocle [8], and Monocle2 [12]—include a built-in functionality for identifying DE genes after pseudotime inference. Their common approach is to use the generalized additive model (GAM) [13–15] to fit each gene's expression level in a cell as a smooth-curve function of the cell's inferred pseudotime. However, these built-in methods for DE gene identification are restricted as an add-on and downstream step of the pseudotime inference method in the same software package, and they cannot take external, user-provided pseudotime as input. Therefore, if users would like to use a new pseudotime inference method, they cannot use these built-in DE methods.

To our knowledge, only two DE gene identification methods can take any user-provided pseudotime. The first and state-of-the-art one is tradeSeq, which uses the negative binomial generalized additive model (NB-GAM) to model the relationship between each gene's expression in a cell and the cell's pseudotime [16]. Its p-value calculation is based on a chi-squared distribution, an inaccurate approximation to the null distribution. As a result, its p-values lack the correct probability interpretation. This issue is reflected in the tradeSeq paper, where the authors acknowledged, "Rather than attaching strong probabilistic interpretations to the p-values (which, as in most RNA-seq applications, would involve a variety of hard-to-verify assumptions and would not necessarily add much value to the analysis), we view the p-values simply as useful numerical summaries for ranking the genes for further inspection." Hence, the uncalibrated p-values of tradeSeq cannot be used for p-value-based statistical procedures such as the type I error control and the false discovery rate (FDR) control. The second method is Monocle3, better known as a pseudotime inference method [17], yet it also allows DE gene identification based on user-provided cell covariates via regression analysis. For clarity, we refer to the pseudotime inference and differential expression functionalities in Monocle3 as "Monocle3-PI" and "Monocle3-DE," respectively. Monocle3-DE uses the generalized linear model (GLM) to identify DE genes for a user-provided covariate, e.g., pseudotime. However, GLM is more restrictive than GAM in that GLM assumes the logarithmic transformation of a gene's expected read count in a cell is a strictly linear function of the cell's pseudotime, while this assumption does not hold for many genes [18]. Hence, Monocle3-DE would miss those complex relationships between gene expression and pseudotime that do not satisfy its GLM assumption. In other words, Monocle3-DE's restrictive GLM assumption impairs its power in identifying DE genes.

Besides the scRNA-seq methods we mentioned above, there are methods developed for identifying physical-time-varying DE genes from bulk RNA-seq time-course data. Among those methods, the ones allowing for continuous time can in principle be used to identify DE genes along pseudotime. Two examples of such methods are NBAMSeq [19] and ImpulseDE2 [20]. NBAMSeq is similar to tradeSeq in the use of NB-GAM, but it uses the Bayesian shrinkage method in DESeq2 [21] to estimate gene variances, while tradeSeq does not. ImpulseDE2 [20], a method favorably rated in a benchmark study for bulk RNA-seq data [22], models gene differential expression by a unique "impulse" model. A later study modified ImpulseDE2 to identify DE genes along pseudotime from scRNA-seq data [16]. However, the performance of NBAMSeq and ImpulseDE2 on scRNA-seq data lacks benchmarking. Loosely related, many methods can identify DE genes between discrete cell clusters, groups, or conditions [23–27]; however, these methods are inapplicable to finding DE genes along continuous pseudotime.

More importantly, the existing methods that identify DE genes along pseudotime have a common limitation: they ignore the uncertainty of inferred cell pseudotime, which they consider as one fixed value per cell. This issue arises from the fact that most pseudotime inference methods only return point estimates of cell pseudotime without uncertainty quantification (i.e., every cell only receives an inferred pseudotime without a standard error), with few exceptions [28, 29]. Hence, downstream DE gene identification methods treat these point estimates as fixed and ignore their uncertainty. However, this ignorance of uncertainty would result in invalid p-values, leading to either failed FDR control or power loss. This critical problem has been noted in several pseudotime inference method papers [10, 11, 28] and in the tradeSeq paper [16], yet it remains an open challenge to our knowledge.

Motivated by the ill-posed p-value issue of existing pseudotime-based differential expression methods, we propose PseudotimeDE, the first method that adapts to all pseudotime inference methods, takes into account the random nature of inferred pseudotime, and outputs well-calibrated p-values. PseudotimeDE uses subsampling to estimate pseudotime inference uncertainty and propagates the uncertainty to its statistical test for DE gene identification. As the most notable advantage of PseudotimeDE over existing methods, PseudotimeDE's well-calibrated p-values ensures the reliability of FDR control and other downstream analyses, as well as avoiding unnecessary power loss due to overly-conservative p-values.

Results

Overview of the PseudotimeDE method

The statistical method of PseudotimeDE consists of four major steps: subsampling, pseudotime inference, model fitting, and hypothesis testing (Fig. 1). The first two steps are performed at the cell level and include all informative genes (whose selection depends on the pseudotime inference method, e.g., Slingshot and Monocle3-PI), while the last two steps are performed on every gene that is potentially DE.

1. In the subsampling step, PseudotimeDE subsamples 80% of cells from the original dataset to capture the uncertainty of pseudotime inference, the same technique as used in [9, 11, 30].

2. In the pseudotime inference step, PseudotimeDE applies a user-specified pseudotime inference method to the original dataset and each subsample, so that every cell receives its inferred pseudotime in the original dataset and all the subsamples that include it. To construct null cases where genes are non-DE for later hypothesis testing, PseudotimeDE permutes the inferred pseudotime in each subsample, independent of other subsamples.

3. In the model fitting step, PseudotimeDE fits NB-GAM or zero-inflated negative binomial GAM (ZINB-GAM) to every gene in the original dataset to obtain a test statistic that indicates the effect size of the inferred pseudotime on the gene's expression.

4. In the hypothesis testing step, for every gene, Pseudotime fits the same model used for the original dataset to the permuted subsamples to obtain approximate null values of the gene's test statistic (the null values are approximate because the subsamples do not have the same number of cells as in the original dataset). To save the number of subsamples needed and to improve the p-value resolution, Pseudotime fits a Gamma distribution or a mixture of two Gamma distributions to these null values. It subsequently uses the fitted parametric distribution as the approximate null distribution of the test statistic. Finally, PseudotimeDE calculates a right-tail p-value for the gene from the gene's test statistic in the original dataset and the approximate null distribution.

Further detail of PseudotimeDE is described in **Methods**.

Simulations verify that pseudotimeDE outperforms existing methods in the validity of p-values and the identification power

We use a widely-used simulator `dyntoy` [9, 16] to generate four synthetic scRNA-seq datasets, among which three are single-lineage datasets with low-, medium- and high-dispersion levels, and the other is a bifurcation dataset. Since the single-lineage high-dispersion dataset best resembles the real scRNA-seq data (Fig. S1), we use it as our primary case. We apply two pseudotime inference methods—Slingshot and Monocle3-PI—to each synthetic dataset to infer cell pseudotime.

First, we find that PseudotimeDE successfully captures the underlying uncertainty of inferred pseudotime. The first layer—"linear uncertainty"—reflects the randomness of inferred cell pseudotime within a cell lineage (Fig. 2a & c). Fig. 2b & d show the distributions of individual cells' inferred pseudotime by Slingshot and Monocle3-PI, respectively, across 1000 subsampled datasets, confirming that linear uncertainty is specific to pseudotime inference methods. Between the two methods, Monocle3-PI demonstrates greater linear uncertainty. The second layer—"topology uncertainty"—reflects the randomness of lineage construction. The synthetic bifurcation dataset contains two cell lineages. Slingshot correctly constructs the bifurcation topology from the original dataset and the 1000 subsampled datasets. While Monocle3-PI captures the bifurcation topology from the original dataset (Fig. 2e), it fails to capture the topology from over 50% of subsamples (Fig. 2f shows randomly picked 10 subsamples), demonstrating its greater topology uncertainty than Slingshot's.

After confirming pseudotime inference uncertainty, we benchmark PseudotimeDE against four DE gene identification methods: tradeSeq, Monocle3-DE, NBAMSeq, and ImpulseDE2. The first two methods, tradeSeq and Monocle3-DE, are the state-of-the-art for scRNA-seq data analysis and thus serve as the main competitors of PseudotimeDE. In our benchmark, we first evaluate these methods in terms of the validity of their p-values, which should be uniformly distributed between 0 and 1 under the null hypothesis (i.e., a gene is not DE). Our results show that, among the five methods, PseudotimeDE generates the best-calibrated p-values that follow the expected uniform distribution most closely (Fig. 3a & f and Figs. S3–S5a & f in the Supplementary). Among the existing four methods, only Monocle3-DE provides roughly calibrated p-values, while tradeSeq, NBAMSeq, and ImpulseDE2 output p-values that are much deviated from the expected uniform distribution. This observation is confirmed by the Kolmogorov–Smirnov test, which evaluates how closely p-values follow the uniform distribution. Since the identification of DE genes relies on a small p-value cutoff, the smaller p-values are more important than the larger ones. Hence, we re-plot the p-values on $-\log_{10}$ scale to closely examine the calibration of small p-values (Fig. 3b & g and Figs. S3–S5b & g in the Supplementary). Again, PseudotimeDE returns the best-calibrated p-values, while the other four methods generate overly small p-values that would inflate false discoveries. This is reflected in our results: at a target 5% FDR threshold, PseudotimeDE leads to the best FDR control among all methods (Fig. 3c & h and Figs. S3–S5c & h in the Supplementary).

Next, we compare these methods in terms of their ability to distinguish DE genes from non-DE genes, ability measured by the area under the receiver operating characteristic curve (AUROC) values (Fig. 3d & i and Figs. S3–S5d & i in the Supplementary). PseudotimeDE achieves the highest AUROC values. Among the other four methods, tradeSeq and NBAMSeq have slightly lower AUROC values than PseudotimeDE's, and Monocle3-DE and ImpulseDE2 have much lower AUROC values than the other three methods'. The reason is that PseudotimeDE, tradeSeq, and

NBAMSeq all use the flexible model NB-GAM, while Monocle3-DE and ImpulseDE2 use much more restrictive models, which limit their power.

Realizing that the ill-calibrated p-values of the existing four methods invalidate their FDR control, we compare all five methods in terms of their power under an actual 5% false discovery proportion (FDP, defined as the proportion of false discoveries among the discoveries in one synthetic dataset) instead of the nominal 5% FDR. Our results show that PseudotimeDE achieves the highest power on all datasets except for the bifurcation dataset, where PseudotimeDE has slightly lower power than tradeSeq's (Fig. 3e & j and Figs. S3–S5e & j in the Supplementary). These results demonstrate the high power of PseudotimeDE and its effective FDR control, which is lacking in existing methods.

In summary, our simulation results verify that PseudotimeDE outperforms existing methods in terms of generating well-calibrated p-values, which are essential for FDR control, and identifying DE genes with high power. Notably, the two bulk RNA-seq methods, NBAMSeq and ImpulseDE2, yield worse results than the three scRNA-seq methods do. Hence, we only focus on the scRNA-seq methods in the following three real data applications.

Real data example 1: dendritic cells stimulated with lipopolysaccharide

In the first application, we compare PseudotimeDE with tradeSeq and Monocle3-DE on a dataset of mouse dendritic cells (DCs) after stimulation with lipopolysaccharide (LPS, a component of gram-negative bacteria) [31]. In this dataset, gene expression changes are expected to be associated with the immune response process. We first apply Slingshot and Monocle3-PI to this dataset to infer cell pseudotime, and then we input the inferred pseudotime into PseudotimeDE, tradeSeq, and Monocle3-DE for DE gene identification. Consistent with our simulation results, the p-values of tradeSeq are ill-calibrated: their bimodal distributions indicate that they do not follow the uniform distribution under the null hypothesis; instead, many of them are inflated, and this inflation would lead to power loss in DE gene identification (Fig. 4a & e). Indeed, at a nominal Benjamini-Hochberg (BH) adjusted p-value ≤ 0.01 threshold (which corresponds to controlling the FDR $\leq 1\%$ when p-values are valid), tradeSeq identifies the smallest number of DE genes, while PseudotimeDE identifies the most DE genes, followed by Monocle3-DE. Notably, most of the DE genes identified by tradeSeq are also identified by PseudotimeDE (Fig. 4b & f), a result consistent with the over-conservativeness of tradeSeq due to its inflated p-values. Unlike tradeSeq, Monocle3-DE does not exhibit the inflated p-value issue; however, it uses a more restrictive model than PseudotimeDE and tradeSeq do. Hence, we use functional analyses to investigate whether Monocle3-DE misses certain DE genes due to its restrictive modeling. We also investigate whether the additional DE genes found by PseudotimeDE but missed by tradeSeq or Monocle3-DE are biologically meaningful.

Our first strategy is to perform gene ontology (GO) analysis on the DE genes identified by each method and compare the enriched GO terms. We find that more GO terms are enriched (with enrichment p-values < 0.01) in the DE genes identified by PseudotimeDE (Fig. S6a & c), and that the PseudotimeDE-specific GO terms are related to immune responses (Fig. S6b & d; Supplementary Table 1). However, comparing enriched GO terms does not directly reflect the difference of DE genes identified by different methods. Hence, our second strategy is to probe the functions of the DE genes that are uniquely identified by one method in pairwise comparisons of PseudotimeDE vs. tradeSeq and PseudotimeDE vs. Monocle3-DE. We first perform GO analysis on each set of uniquely identified DE genes. For a fair comparison of two methods, we remove the overlapping DE genes found by both methods from the background gene list in GO analysis. Our results show that many more GO terms are enriched (with enrichment p-values < 0.01) in Pseudotime-specific DE genes than in tradeSeq- or Monocle3-DE-specific DE genes (Fig. 4c & g). Moreover, many of those PseudotimeDE-specific GO terms are directly related to the immune responses of DCs to LPS stimulation, including the GO terms “cellular response to lipopolysaccharide” and “defense response to Gram-negative bacterium” (Fig. 4d & h; Supplementary Table 1). To focus more on immune responses, we next perform enrichment analysis using the immunologic signatures (C7) in the Molecular Signatures Database (MSigDB) [32]. Our results show that only PseudotimeDE-specific DE genes have enriched MSigDB C7 terms (with BH adjusted p values < 0.01), while tradeSeq- and Monocle3-DE-specific DE genes have almost no enrichment (Fig. S7a & c). More importantly, many enriched terms in PseudotimeDE-specific DE genes were found by previous studies of DCs stimulated with LPS (see examples in Fig. S7b & d; Supplementary Table 1); this is direct evidence that supports the validity of PseudotimeDE-specific DE genes. For illustration purpose, we visualize the expression levels of some known and novel DE genes identified by PseudotimeDE using UMAP, and clear DE patterns are observed (Figs. S8–S9). In conclusion, our functional analyses verify that PseudotimeDE identifies biologically meaningful DE genes missed by tradeSeq and Monocle3-DE, confirming that PseudotimeDE has high power in addition to its well-calibrated p-values.

Real data example 2: pancreatic beta cell maturation

In the second application, we compare PseudotimeDE with tradeSeq and Monocle3-DE on a dataset of mouse beta cell maturation process [33]. We first apply Slingshot and Monocle3-PI to this dataset to infer cell pseudotime, and then we input the inferred pseudotime into PseudotimeDE, tradeSeq, and Monocle3-DE for DE gene identification. Consistent with previous results, the p-values of tradeSeq follow a bimodal distribution, suggesting that many of them are incorrectly inflated (Fig. 5a & f). At the nominal BH-adjusted p-value ≤ 0.01 level, PseudotimeDE identifies the second most DE genes, fewer than Monocle3-DE’s identified DE genes and much more than tradeSeq’s (Fig. 5b & g). As the numbers of identified DE genes cannot reflect these methods’ performance, we use three approaches to evaluate the DE genes identified by each method.

We first perform GO analysis on each set of uniquely identified DE genes, using the same pairwise comparisons of PseudotimeDE vs. tradeSeq and PseudotimeDE vs. Monocle3-DE as for the LPS-dendritic data. Our results show that more GO terms are enriched (with enrichment p-values < 0.01) in PseudotimeDE-specific DE genes than in tradeSeq- or Monocle3-DE-specific DE genes (Fig. 5c & h). Moreover, many of those PseudotimeDE-specific GO terms are directly related to pancreatic beta cell development, e.g., "positive/negative regulation of Notch signaling pathway" [34] and "endocrine pancreas development" (Fig. 5c & h; Supplementary Table 2). As a complementary result, we also perform GO analysis on the DE genes identified by each method. We find that the GO terms, which are only enriched in the DE genes identified by PseudotimeDE, are related to beta cell development and thus more biologically meaningful than the GO terms that are only enriched in the DE genes identified by tradeSeq or Monocle3-DE (Fig. S10b & d; Supplementary Table 2).

Second, we utilize the DE genes identified from bulk RNA-seq data in the original paper [33] to evaluate the DE gene rankings established by PseudotimeDE, tradeSeq, and Monocle3-DE from scRNA-seq data. Taking the bulk DE genes as a gene set, we perform the gene-set enrichment analysis (GSEA) [32] on all genes' $-\log_{10} p$-values output by PseudotimeDE, tradeSeq, and Monocle3-DE. Among the three methods, PseudotimeDE leads to the highest normalized enrichment score (NES) (Supplementary Table 3), suggesting that the bulk DE genes are most enriched in the top-ranked DE genes found by PseudotimeDE.

Third, we examine a highly credible DE gene *Slc39a10* [33, 35] and a verified non-DE gene *Sst* [33] as representative examples. For *Slc39a10*, both PseudotimeDE and Monocle3-DE yield small p-values ($< 10^{-6}$), while tradeSeq outputs a p-value > 0.1 and thus misses it (Fig. 5e & g). For *Sst*, PseudotimeDE yields the largest p-value (> 0.001), while tradeSeq and Monocle3-DE yield extremely small p-values ($< 10^{-10}$) and thus mistaken it as a DE gene. Hence, PseudotimeDE has the best performance on these two representative genes.

For illustration purpose, we visualize the expression levels of some known and novel DE genes identified by PseudotimeDE using UMAP, and clear DE patterns are observed (Figs. S11–S12).

Real data example 3: bone marrow differentiation

In the third application, we compare PseudotimeDE with tradeSeq and Monocle3-DE on a dataset of mouse bone marrow differentiation [36]. We apply Slingshot with UMAP for dimensionality reduction to infer cell pseudotime as described in the tradeSeq paper [16]. Slingshot constructs the reported bifurcation topology (in the tradeSeq paper) on the original dataset (Fig. 6a), but it infers trifurcation topology, instead of bifurcation topology, on 40% of subsamples (Fig. 6b shows randomly picked ten subsamples). Note that the third lineage consisting of the cell type megakaryocyte (MK) was reported in the Monocle2 paper ([12]), suggesting the observed topology uncertainty may be biologically meaningful.

For a fair comparison, we only make pseudotimeDE use the subsamples with inferred bifurcation topology, because both tradeSeq and Monocle3-DE use the inferred bifurcation topology from the original data to identify DE genes. Consistent with previous results, the tradeSeq p-values follow a bimodal distribution that is unexpected for well-calibrated p-values. At a nominal BH-adjusted p-value ≤ 0.01 threshold, the three methods identify highly similar DE genes (Fig. 6e & g). For instance, PseudotimeDE and tradeSeq share about 80% of their identified DE genes (Jaccard index). From the few method-specific DE genes, functional analyses cannot indicate which method performs better. Therefore, we use GSEA instead to evaluate methods' p-values. Surprisingly, although the three methods identify highly similar DE genes, their p-values lead to vastly different GSEA results. At the $q < 0.25$ level, PseudotimeDE and Monocle3-DE yield hundreds of enriched gene sets, while tradeSeq only yields a few or no enriched gene sets (Fig. 6f & h; Supplementary Table 3). This result indicates that, besides the ranking of p-values, the nominal values of p-values are also crucial for downstream analysis such as GSEA. Hence, the well-calibrated p-values make PseudotimeDE superior to existing methods for DE gene identification and downstream analyses.

Real data example 4: natural killer T cell subtypes

In the fourth application, we compare PseudotimeDE with tradeSeq and Monocle3-DE on a dataset of natural killer T cell (NKT cell) subtypes [37]. We apply Slingshot with PCA for dimensionality reduction to infer cell pseudotime and construct the trifurcation topology (Fig. 7a) reported in the original study. We apply the three DE methods to identify DE genes in each of the three lineages. Consistent with the previous results, the p-values of tradeSeq follow a bimodal distribution, suggesting that many of them are incorrectly inflated (Fig. 7b).

For validation purpose, we utilize the lineage-specific DE genes identified from bulk RNA-seq data in the original study [37] to evaluate the DE gene rankings established by PseudotimeDE, tradeSeq, and Monocle3-DE from scRNA-seq data. Specifically, we perform the GSEA analysis using the bulk DE gene sets in the same way as for the pancreatic beta cell maturation dataset. The GSEA analysis shows that PseudotimeDE's p-values best agree with the lineage-specific DE genes from bulk data and thus most distinguish the three lineages. For example, for the NKT1 lineage, PseudotimeDE's small p-values are exclusively enriched in the "NKT1 bulk" gene set, while tradeSeq and Monocle3-DE have small p-values enriched in at least two lineage-specific DE gene sets (Fig. 7c). This result confirms that, compared with the DE genes identified by the other two DE methods, the top DE genes identified by PseudotimeDE are more biologically meaningful.

Real data example 5: cell cycle phases

In the fifth application, we compare PseudotimeDE with tradeSeq and Monocle3-DE on a dataset of human induced pluripotent stem cells (iPSCs) measured with cell cycle phases (FUCCI labels) [38]. The original study has reported 101 cyclic genes whose expression levels have large proportions of variance explained (PVE) explained by cells' FUCCI labels [38]; that is, cells' FUCCI labels are regarded as the predictor, a gene's expression levels in the same cells are regarded as the response, and a PVE is calculated from a nonparametric smoothing fit; hence, the larger the PVE, the better the gene's expression levels can be predicted by the cell cycle phases. The original study has also developed an R package `peco` to infer cell cycle phases from scRNA-seq data.

In our study, we first construct a benchmark dataset by treating the 101 cyclic genes as true DE genes and using the same genes with expression levels randomly shuffled across cells as the true non-DE genes; hence, our positive and negative sets both contain 101 genes. Then we apply the R package $\texttt{peco}$ to this dataset to infer each cell's cycle phase, which is equivalent to pseudotime; that is, we use $\texttt{peco}$ as the pseudotime inference method. Finally, we apply the three DE methods.

Our results show that, for the true non-DE genes, only PseudotimeDE generates valid p-values that approximately follow the Uniform$[0, 1]$ distribution (Fig. 8a & c). For the true DE genes, PseudotimeDE's $(-\log_{10}$ transformed) p-values, one per gene, have the highest correlation with these genes' PVE, indicating that PseudotimeDE successfully identifies the top DE genes as those with the strongest cyclic trends (Fig. 8b). PseudotimeDE also yields successful FDR control, the highest AUROC value, and the highest power, among the three DE methods (Fig. 8d, e and f). Therefore, we conclude that PseudotimeDE outperforms tradeSeq and Monocle3-DE in identifying cell cycle-related genes from this iPSC scRNA-seq dataset.

PseudotimeDE allows users to inspect the uncertainty of inferred cell pseudotime

Besides the identification of DE genes, PseudotimeDE offers functionality for inspecting the uncertainty of pseudotime inference via its intermediate subsampling step. In Figs. 2e & f and 6a & b, PseudotimeDE reveals the uncertainty of inferred cell lineages. Users usually want to fix the lineage topology, e.g., a bifurcation topology, for downstream analysis; however, the topology can vary across subsamples. Hence, we recommend that users check if the inferred topology from the original data can also be inferred from more than half of the subsamples. If not, users may consider using another pseudotime inference method with less uncertainty or adding additional constraints on inferred topology. Otherwise, the great uncertainty of inferred cell lineage would impair the reliability of downstream analyses.

Next, conditioning on a given lineage topology, PseudotimeDE allows users to visualize the uncertainty of pseudotime within a lineage (Fig. 2a–d), also guiding the choice of pseudotime inference methods in terms of uncertainty.

Computational time

The only feasible way to accommodate all pseudotime inference methods and to account for their uncertainty in DE gene identification is to use subsampling and permutation, the approach taken by PseudotimeDE. However, a common concern of permutation-based methods is that they are computationally intensive. Admittedly, PseudotimeDE is slower than existing non-permutation-based methods, but its computational time is nevertheless acceptable to server users. With 24 cores (Intel ”Cascade Lake” CPU), 36 GB RAM and 1000 subsamples, PseudotimeDE takes 3–8 hours to analyze each of the three scRNA-seq datasets in our study. Specifically, the LPS-dendritic cell dataset (4016 genes, 390 cells) takes 3 hours, the pancreatic beta cell maturation dataset (6121 genes, 497 cells) takes 3.5 hours, and the bone marrow dataset (3004 genes, 2660 cells, two lineages) takes 8 hours. The computational time is proportional to the number of genes, the number of lineages, and the number of subsamples. Of course, it is inversely proportional to the number of available cores.

To reduce the computational time of PseudotimeDE, users have two options. First, they may reduce the number of genes to be tested. For instance, lowly expressed genes, such as those with more than 90% of zero counts, are recommended to be filtered out because they are usually of less interest to biologists. Second, they may reduce the number of subsamples. Due to its parametric estimation of the null distribution of the test statistic, PseudotimeDE does not require an enormous number of subsamples. We find that PseudotimeDE with only 100 subsamples generates similar p-values to those based on 1000 subsamples (Fig. S20). If using 100 subsmaples, the computational time is within 0.5 hour.

In an undesirable scenario that computational resources are too limited, users have to abandon the consideration of pseudotime uncertainty and treat inferred pseudotime as fixed. Then they do not need the subsampling procedure, and PseudotimeDE will calculate p-values from the asymptotic null distribution of the test statistic [39], with a short computational time as other non-permutation-based methods'.

Discussion

We propose a statistical method PseudotimeDE to identify DE genes along inferred cell pseudotime. PseudotimeDE focuses on generating well-calibrated p-values while taking into account the randomness of inferred pseudotime. To achieve these goals, PseudotimeDE first uses subsampling to estimate the uncertainty of pseudotime. Second, PseudotimeDE fits the NB-GAM or ZINB-GAM to both the original dataset and the permuted subsampled datasets to calculate the test statistic and its approximate null values. Next, PseudotimeDE fits a parametric distribution to estimate the approximate null distribution of the test statistic. Finally, PseudotimeDE calculates p-values with a high resolution. PseudotimeDE is flexible to accommodate cell pseudotime inferred in a standard format by any method. Its use of NB-GAM and ZINB-GAM allows it to capture diverse gene expression dynamics and to accommodate undesirable zero inflation in data.

Comprehensive studies on simulated and real data confirm that PseudotimeDE yields better FDR control and higher power than four existing methods (tradeSeq, Monocle3-DE, NBAMSeq, and ImpulseDE2) do. On simulation data, PseudotimeDE generates well-calibrated p-values that follow the uniform distribution under the null hypothesis, while existing methods except Monocle3-DE have p-values violating the uniform assumption. Well-calibrated p-values guarantee the valid FDR control of PseudotimeDE. Moreover, thanks to its use of flexible models NB-GAM and ZINB-GAM, PseudotimeDE has higher power than Monocle3-DE, which uses a more restrictive model GLM and thus has less power. PseudotimeDE also outperforms the other three methods–tradeSeq, NBAMSeq, and ImpulseDE2—that generate ill-calibrated p-values in terms of power. On three real scRNA-seq datasets, the DE genes uniquely identified by PseudotimeDE embrace better biological interpretability revealed by functional analyses, and the p-values of PseudotimeDE lead to more significant GSEA results.

An interesting and open question is what pseudotime inference method works the best with PseudotimeDE. While we observe that PseudotimeDE has higher power with Slingshot than with Monocle3-PI in simulation studies, we realize that the reason may be associated with the simulation design (e.g., the lineage structures), and thus we cannot draw a conclusion from this observation. Due to the diversity of biological systems and the complexity of pseudotime inference [9], we decide to leave the choice of pseudotime inference methods open to users, and this is the advantage of PseudotimeDE being flexible to accommodate inferred pseudotime by any methods. In practice, we encourage users to try popular pseudotime inference methods and use PseudotimeDE as a downstream step to identify DE genes, so that they can analyze the identification results and decide which pseudotime inference method is more appropriate for their dataset.

The zero inflation, or “dropout” issue, remains perplexing and controversial in the single-cell field [40–42]. The controversy is regarding whether excess zeros that cannot be explained by Poisson or negative binomial distributions are biological meaningful or not. Facing this controversy, we provide two models in PseudotimeDE: NB-GAM and ZINB-GAM, the former treating excess zeros as biologically meaningful and the latter not. Specifically, the negative binomial distribution in NB-GAM is fitted to all gene expression counts including excess zeros, while the fitting of the negative distribution in ZINB-GAM excludes excess zeros, which ZINB-GAM implicitly treats as non-biological zeros. PseudotimeDE allows the choice between the two models to be user specified or data driven. From our data analysis, we realize that the choice often requires biological knowledge of the dataset to be analyzed. Specifically, on the LPS-dendritic cell dataset and pancreatic beta cell maturation dataset, we observe that ZINB-GAM leads to power loss: some potential DE genes cannot be identified by ZINB-GAM because zero counts contain useful information (Figs. S15–S18). Our observation is consistent with another recent study [42], whose authors observed that “zero-inflated models lead to higher false-negative rates than identical non-zero-inflated models.” Hence, our real data analysis results are based on NB-GAM. However, realizing the complexity of biological systems and scRNA-seq protocols, we leave the choice between NB-GAM and ZINB-GAM as an option for users of PseudotimeDE, and we encourage users to plot their known DE genes as in Figs. S15–S18 to decide which of NB-GAM and ZINB-GAM better captures the gene expression dynamics of their interest.

The current implementation of PseudotimeDE is restricted to identifying the DE genes that have expression changes within a cell lineage. While methods including GPfates [43], Monocle2 BEAM [44], and tradeSeq can test whether a gene’s expression change is associated with a branching event leading to two lineages, they do not consider the uncertainty of lineage inference. How to account for such topology uncertainty is a challenging open question, as we have seen in

Figs. 2f and 6b that the inferred lineage may vary from a bifurcation topology to a trifurcation topology on different subsets of cells. A possible direction is to use the selective inference [45, 46], and we will leave the investigation of this question to future research. Due to this topology uncertainty issue, PseudotimeDE is most suitable for single-cell gene expression data that contain only one cell lineage (including cyclic data) or a small number of well separated cell lineages (e.g., bifurcation). The reason is that these data can maintain stable inferred cell topology after subsampling, an essential requirement of PseudotimeDE. That said, PseudotimeDE is not designed for data with many equivocal cell lineages or a complex cell hierarchy, the data that cannot maintain stable inferred cell topology across subsamples, because for such data, it is difficult to find one-to-one matches between cell lineages inferred from a subsample and those inferred from the original data. Then, a practical solution for such data is to first define a cell lineage of interest and then apply PseudotimeDE to only the cells assigned to this lineage.

There are other open questions to be explored. An important question is: when do we want to identify DE genes along pseudotime? As we have shown in the **Results** section, inferred pseudotime can be highly uncertain. As biologists often sequence cells at multiple physical time points if they want to investigate a biological process, a straightforward analysis is to identify the DE genes that have expression changes across the physical time points. Then we have two definitions of DE genes: the genes whose expression changes across pseudotime vs. physical time. Understanding which definition is more biologically relevant is an open question. Another question is whether it is possible to integrate pseudotime with physical time to identify biologically relevant DE genes. Answering either question requires a statistical formulation that is directly connected to a biological question.

Another question is how to explore gene-gene correlations along cell pseudotime. Current DE methods only detect marginal gene expression changes but ignore gene-gene correlations. It remains unclear whether gene-gene correlations are stable or varying along cell pseudotime. Hence, a new statistical method to detect gene-gene correlation changes along inferred pseudotime may offer new biological insights into gene co-expression and regulation at the single-cell resolution.

Methods

PseudotimeDE methodology

Here we describe the methodological detail of PseudotimeDE, a pseudotime-based differential expression testing method. As an overview, PseudotimeDE takes a scRNA-seq count matrix and a pseudotime inference method as input, estimates the uncertainty of pseudotime, performs a differential expression test, and returns a p-value of each gene. The core of PseudotimeDE is to obtain a valid null distribution of the DE gene test statistic so that the p-values are well-calibrated.

Mathematical notations

We denote by $\mathbf{Y} = (Y_{ij})$ an $n \times m$ gene expression count matrix, whose rows and columns correspond to n cells and m genes, respectively; that is, Y_{ij} is the read count of gene j in cell i. Taking $\mathbf{Y}$ as input, a pseudotime inference method would return a pseudotime vector $\mathbf{T} = (T_1, \ldots, T_i, \ldots, T_n)^\mathsf{T}$, where $T_i \in [0, 1]$ denotes the normalized inferred pseudotime of cell i (i.e., the cells with the smallest and largest pseudotime have $T_i = 0$ and 1, respectively; normalization is used for visualization simplicity). Note that T_i is a random variable due to the random-sampling nature of the n cells and the possible uncertainty introduced by the pseudotime inference method.

Uncertainty estimation

To estimate the uncertainty of pseudotime $\mathbf{T}$, we subsample 80% cells (columns) in $\mathbf{Y}$ for B times. Although there are some theoretical results about the optimal subsample size [47], they do not apply to our problem setting. Hence, we simply choose 80% because it is widely used [48, 49], similar to the popularity of 5-fold cross validation in machine learning [50]. Simulation results also supports that 80% is a reasonable choice, and PseudotimeDE is robust to various subsampling proportions (Fig. S22; see Supplementary for detail). It is worth noting that the bootstrap technique is inapplicable for our problem because it leads to repeated sampling of the same cell, causing issues with some pseudotime inference methods such as Monocle2. If the cells have pre-defined groups (i.e., cell types), we use the stratified sampling by first subsampling 80% cells within each group and then combining these within-group subsamples into one subsample. By default, we set $B = 1000$. For each subsample $\mathbf{Y}^b = (Y_{ij}^b)$, an $n' \times m$ matrix where $n' = \lfloor .8n \rfloor$, we perform pseudotime inference with the same parameters used for the original dataset $\mathbf{Y}$. As a result, we obtain B subsample-based realizations of pseudotime $\mathbf{T}$: $\left\{\mathbf{T}^1, \cdots, \mathbf{T}^b, \cdots, \mathbf{T}^B\right\}$, where $\mathbf{T}^b \in [0, 1]^{n'}$, and each cell appears in approximately 80% of these B realizations. Note that we have to apply pseudotime inference to each subsample before permutation to account for pseudotime inference uncertainty; otherwise, if each subsample's pseudotime is just a subsample of all cells' pseudotime, we are essentially treating all cells' pseudotime as fixed, and the uncertainty in pseudotime inference is ignored. Here is the mathematical explanation. Given that we have n cells with inferred pseudotime as $T_1, \ldots, T_n$, if we use direct subsampling, then in the b-th subsampling, the subsampled n' cells' pseudotime is just a size-n' subsample of $\{T_1, \ldots, T_n\}$. Instead, in PseudotimeDE, the subsampled n' cells' inferred pseudotime $T_1^b, \ldots, T_{n'}^b$ may be n' values that are not in $T_1, \ldots, T_n$. In other words, the uncertainty in pseudotime inference is reflected in $T_1^b, \ldots, T_{n'}^b$.

PseudotimeDE model

We use the negative binomial–generalized additive model (NB-GAM) as the baseline model to describe the relationship between every gene's expression in a cell and the cell's pseudotime. For gene j ($j = 1, \ldots, m$), its expression Y_{ij} in cell i and the pseudotime T_i of cell i ($i = 1, \ldots, n$) are assumed to follow

$$\begin{cases} Y_{ij} \sim \mathrm{NB}(\mu_{ij}, \phi_j)\,, \\ \log(\mu_{ij}) = \beta_{j0} + f_j(T_i)\,, \end{cases}$$

where $\mathrm{NB}(\mu_{ij}, \phi_j)$ denotes the negative binomial distribution with mean μ_{ij} and dispersion ϕ_j, and $f_j(T_i) = \sum_{k=1}^{K} b_k(T_i)\beta_{jk}$ is a cubic spline function. The number of knots k is predefined as 6 and usually has little effect on results [51]. For gene j, PseudotimeDE fits the NB-GAM to $(Y_{1j}, \ldots, Y_{nj})^{\mathsf{T}}$ and $\mathbf{T} = (T_1, \ldots, T_n)^{\mathsf{T}}$ using the R package `mgcv` (version 1.8.31), which estimates model parameters by penalized-restricted maximum likelihood estimation.

To account for excess zeros in scRNA-seq data that may not be explained by the NB-GAM, we introduce a hidden variable Z_{ij} to indicate the "dropout" event of gene j in cell i, and the resulting model is called the zero-inflated negative binomial–generalized additive model (ZINB-GAM):

$$\begin{cases} Z_{ij} \sim \mathrm{Ber}(p_{ij})\,, \\ Y_{ij}|Z_{ij} \sim Z_{ij} \cdot \mathrm{NB}(\mu_{ij}, \phi_j) + (1 - Z_{ij}) \cdot 0\,, \\ \log(\mu_{ij}) = \beta_{j0} + f_j(T_i)\,, \\ \mathrm{logit}(p_{ij}) = \alpha_{j0} + \alpha_{j1}\log(\mu_{ij})\,. \end{cases}$$

For gene j, PseudotimeDE fits the ZINB-GAM to $(Y_{1j}, \ldots, Y_{nj})^{\mathsf{T}}$ and $\mathbf{T} = (T_1, \ldots, T_n)^{\mathsf{T}}$ using the expectation-maximization (EM) algorithm, which is partially based on R package `zigam` [52]. To use PseudotimeDE, users can specify whether to use the ZINB-GAM or NB-GAM. If users do not provide a specification, PseudotimeDE will automatically choose between the two models for each gene by the Akaike information criterion (AIC). By default, PseudotimeDE uses NB-GAM unless the AIC of ZINB-GAM exceeds the AIC of NB-GAM by at least 10, a threshold suggested by [53].

Statistical test and p-value calculation

To test if gene j is DE along cell pseudotime, PseudotimeDE defines the null and alternative hypotheses as

$$H_0 : f_j(\cdot) = \mathbf{0} \quad \text{vs.} \quad H_1 : f_j(\cdot) \neq \mathbf{0}$$

We denote the estimate of $(f_j(T_1), \ldots, f_j(T_n))^{\mathsf{T}}$ by $\hat{\boldsymbol{f}}_j$, whose estimated covariance matrix (of dimensions $n \times n$) is denoted by $\widehat{\mathbf{V}}_{f_j}$. Then the test statistic is

$$S_j = \hat{\boldsymbol{f}}_j^{\mathsf{T}} \widehat{\mathbf{V}}_{f_j}^{r-} \hat{\boldsymbol{f}}_j\,,$$

where $\widehat{\mathbf{V}}_{f_j}^{r-}$ is the rank-r pseudoinverse of $\widehat{\mathbf{V}}_{f_j}$, where r is determined in the way described in [39]. When the T_i's are fixed, the asymptotic null distribution of S_j is described in [39], and the p-value can be calculated by the R package `mgcv`.

A key novelty of PseudotimeDE is its accounting for the uncertainty of inferred pseudotime. When the T_i's are random, the asymptotic null distribution of S_j given that T_i's are fixed [39] and the

p-value calculation in the R package `mgcv` no longer apply. To address this issue and estimate the null distribution, PseudotimeDE uses the following permutation procedure: (1) PseudotimeDE randomly permutes each subsample-based realization $\mathbf{T}^b = (T_1^b, \ldots, T_{n'}^b)^\mathsf{T}$ into $\mathbf{T}^{*b} = (T_1^{*b}, \ldots, T_{n'}^{*b})^\mathsf{T}$; (2) PseudotimeDE fits the above model to $(Y_{1j}^b, \ldots, Y_{n'j}^b)^\mathsf{T}$ and $\mathbf{T}^{*b}$, and calculates the test statistic S_j's value as s_j^b using the R package `mgcv`; (3) PseudotimeDE performs (1) and (2) for $b = 1, \ldots, B$ and collects the resulting $\{s_j^1, \ldots, s_j^B\}$ as the null values of the test statistic S_j.

Then PseudotimeDE estimates the null distribution of S_j in two ways. Based on the estimated null distribution in either way and the observed test statistic value s_j, which is calculated from the original dataset by the R package `mgcv`, PseudotimeDE calculates a p-value for gene j.

1. **Empirical estimate.** PseudotimeDE uses the empirical distribution of $\{s_j^1, \ldots, s_j^B\}$ as the estimated null distribution. Following the suggestion in [54], PseudotimeDE calculates the p-value of gene j as

$$p_j^{\text{emp}} = \frac{\sum_{b=1}^B \mathbb{I}(s_j^b \geq s_j) + 1}{B + 1} \,,$$

where $\mathbb{I}(\cdot)$ is the indicator function. We refer to this p-value as the "empirical p-value."

2. **Parametric estimate.** The resolution of p_j^{emp} depends on the number of permutations B, because the smallest value p_j^{emp} may take is $1/(B+1)$. Although users often cannot afford a too large B due to limited computational resources, they still desire a high resolution of p-values to control the FDR to a small value (e.g., 5%) when the number of tests (i.e., the number of genes in DE gene identification) is large. To increase the resolution of p-values, PseudotimeDE fits a parametric distribution to $\{s_j^1, \ldots, s_j^B\}$ and uses the fitted distribution as the estimated null distribution. Driven by the empirical distribution of $\{s_j^1, \ldots, s_j^B\}$, PseudotimeDE considers two parametric distributions: (1) a gamma distribution $\Gamma(\alpha, \beta)$ with $\alpha, \beta > 0$ and (2) a two-component gamma mixture model $\gamma\Gamma(\alpha_1, \beta_1) + (1-\gamma)\Gamma(\alpha_2, \beta_2)$ with $0 < \gamma < 1$ and $\alpha_1, \beta_1, \alpha_2, \beta_2 > 0$. After fitting both distributions to $\{s_j^1, \ldots, s_j^B\}$ using the maximum likelihood estimation (gamma distribution fit by the R package `fitdistrplus` (version 1.0.14) [55] and gamma mixture model fit by the R package `mixtools` (version 5.4.5) [56]), PseudotimeDE chooses between the two fitted distributions by performing the likelihood ratio test (LRT) with 3 degrees of freedom (i.e., difference in the numbers of parameters between the two distributions). If the LRT p-value is less or equal than 0.01, PseudotimeDE uses the fitted two-component gamma mixture model as the parametric estimate of the null distribution of S_j; otherwise, PseudotimeDE uses the fitted gamma distribution. The Anderson-Darling goodness-of-fit test verifies that such a parametric approach fits the empirical distributions well (Fig. S21; see Supplementary for detail). Denoting the cumulative distribution function of the parametrically estimated null distribution by $\hat{F}_j(\cdot)$, PseudotimeDE calculates the p-value of gene j as

$$p_j^{\text{param}} = 1 - \hat{F}_j(s_j) \,,$$

where is referred to as the "parametric p-value."

PseudotimeDE outputs both p_j^{emp} and p_j^{param} for gene j, $j = 1, \ldots, m$. Empirical evidence shows that parametric p-values agree with empirical p-values well across the $[0, 1]$ interval (Fig. S19). All the findings in the Results section are based on $p_1^{\text{param}}, \ldots, p_m^{\text{param}}$ due to their higher resolution.

Pseudotime inference methods

We apply two state-of-the-art methods, Slingshot and Monocle3-PI, to inferring the cell pseudotime of each dataset. For single-lineage data, we specify the start cluster in Slingshot and the start node in Monocle 3. For bifurcation/trifurcation data, we specify the start cluster/node and the end clusters/nodes in Slingshot/Monocle3-PI. By default, the dimensionality reduction methods used for pseudotime inference are PCA and UMAP for Slingshot and Monocle3-PI, respectively. The R Bioconductor package `slingshot` (version 1.4.0) and the R package `monocle3` (version 0.2.0) are used.

DE analysis methods

We compare PseudotimeDE with four existing methods for identifying DE genes along pseudotime/time-course from scRNA-seq data (tradeSeq and Monocle3-DE) or bulk RNA-seq data (ImpulseDE2 and NBAMSeq). All these methods take a count matrix $\mathbf{Y}$ and a pseudotime vector $\mathbf{T}$ as input, and they return a p-value for each gene. For tradeSeq, we use the functions `fitGAM` and `associationTest` (`https://statomics.github.io/tradeSeq/articles/tradeSeq.html`). The number of knots parameter K in tradeSeq is chosen by 100 random genes based on the tradeSeq vignette. For Monocle3-DE, we use the function `fit_models` (`https://cole-trapnell-lab.github.io/monocle3/docs/differential/`). Since ImpulseDE2 cannot be applied to scRNA-seq data directly, we follow the modified implementation of ImpulseDE2 in the tradeSeq paper (`https://github.com/statOmics/tradeSeqPaper`). The R Bioconductor packages `tradeSeq` (version 1.3.15), `monocle3` (version 0.2.0), `ImpulseDE2` (version 1.10.0), and `NBAMSeq` (version 1.10.0) are used.

Functional (gene ontology and gene-set enrichment) analyses

We use the R package `topGO` (version 2.38.1) [57] to perform the gene-ontology (GO) enrichment analysis on identified DE genes. We use the R package `clusterProfiler` (version 3.14.3) [58] to perform the gene-set enrichment analysis (GSEA) analysis on ranked gene lists, where genes in each list are ranked by their ranking sores defined as $-\log_{10}$ transformed p-values (the gene with the smallest p-value is ranked the top); p-values that are exactly zeros are replaced by one-tenth of the smallest none-zero p-value. If unspecified, the GO terms are "biological process (BP)" terms.

Simulation study

We use the R package `dyntoy` (0.9.9) to generate single-lineage data and bifurcation data. For single-lineage data, we generate three datasets with increasing dispersion levels (low dispersion, medium dispersion, and high dispersion). Each single-lineage dataset consists of 500 cells and 5000 genes (with 20% as DE genes). For bifurcation data, we use the medium dispersion level. The bifurcation dataset consists of 750 cells and 5000 genes (with 20% as DE genes).

Case studies

LPS-dendritic cell dataset: this Smart-seq dataset contains primary mouse dendritic cells (DCs) stimulated with lipopolysaccharide (LPS) [31], available at Gene Expression Omnibus (GEO) under accession ID GSE45719. In our analysis, we use the the cells from 1h, 2h, 4h, and 6h in the pre-processed data from the study that benchmarked pseudotime inference methods [9]. After the genes with $> 90\%$ zeros are removed, the final dataset consists of 4016 genes and 390 cells, which are expected to be in a single lineage. When applying tradeSeq, we use the recommended ZINB-WaVE + tradeSeq procedure to account for potential zero-inflation. The R Bioconductor package `zinbwave` (version 1.8.0) is used.

Pancreatic beta cell maturation dataset: this Smart-seq2 dataset measures the maturation process of mouse pancreatic beta cells [33], available at GEO under accession ID GSE87375. We use the cells from cell type "beta" in the pre-processed data from the study that benchmarked pseudotime inference methods [9]. After the genes with $> 90\%$ zeros are removed, the final dataset consists of 6121 genes and 497 cells, which are expected to be in a single lineage. When applying tradeSeq, we use the recommended ZINB-WaVE + tradeSeq procedure to account for potential zero-inflation. The R Bioconductor package `zinbwave` (version 1.8.0) is used.

Mouse bone marrow dataset: this MARS-seq dataset contains myeloid progenitors in mouse bone marrow [36], available at GEO under accession ID GSE72859. We use the pre-processed data provided by the tradeSeq vignette. After the genes with $> 90\%$ zeros are removed, the final dataset consists of 3004 genes and 2660 cells. We follow the procedure of combining UMAP and Slingshot to infer pseudotime as described in tradeSeq paper [16]

Natural killer T cell dataset: this Smart-seq2 dataset measures four natural killer T cell (NKT cell) subtypes in mouse [37], available at GEO under accession ID GSE74597. We use the pre-processed data from the study that benchmarked pseudotime inference methods [9]. After the genes with $> 90\%$ zeros are removed, the final dataset consists of 5270 genes and 197 cells, which are expected to have three lineages. We use PCA + Slingshot to infer the pseudotime. When applying tradeSeq, we use the recommended ZINB-WaVE + tradeSeq procedure to account for potential zero-inflation. The R Bioconductor package `zinbwave` (version 1.8.0) is used.

Cell cycle phase dataset: this Fluidigm protocol dataset measures human induced pluripotent stem cells (iPSCs) [38]. The iPSCs are FUCCI-expressing so that their cell cycle phases can be tracked. The authors also developed an R package `peco` for predicting cell cycle phases from single-cell gene expression data. We use the example dataset provided by `peco`, which consists of 101 known cell cycle-related genes (DE genes). To construct null cases, we randomly shuffle the 101 DE genes' expression levels across cells to create 101 non-DE genes. The final dataset consists of 202 genes and 888 cells. We use the R package `peco` (version 1.1.21) to infer cell cycle phases.

Software and code

The R package `PseudotimeDE` is available at `https://github.com/SONGDONGYUAN1994/PseudotimeDE`.

Availability of data and materials

To be added to `zenodo`.

Acknowledgements

The authors would like to thank Dr. Kelly Street for his suggestions on the usage of Slingshot and uncertainty estimate. The authors would like to thank Mr. Huy Nguyen for his assistance in simulation. The authors also appreciate the comments and feedback from the members of the Junction of Statistics and Biology at UCLA (http://jsb.ucla.edu).

Funding

This work was supported by the following grants: National Science Foundation DBI-1846216, NIH/NIGMS R01GM120507, Johnson & Johnson WiSTEM2D Award, Sloan Research Fellowship, and UCLA David Geffen School of Medicine W.M. Keck Foundation Junior Faculty Award (to J.J.L.).

Author information

To be added.

Ethics declarations

Ethics approval and consent to participate

None.

Consent for publication

Not applicable for this study.

Competing interests

None.

References

[1] Ashraful Haque, Jessica Engel, Sarah A Teichmann, and Tapio Lönnberg. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome medicine*, 9(1):1–12, 2017.

[2] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.

[3] Efthymia Papalexi and Rahul Satija. Single-cell rna sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, 18(1):35, 2018.

[4] Sophie Tritschler, Maren Büttner, David S Fischer, Marius Lange, Volker Bergen, Heiko Lickert, and Fabian J Theis. Concepts and limitations for learning developmental trajectories from single cell genomics. *Development*, 146(12):dev170506, 2019.

[5] Adam P Croft, Joana Campos, Kathrin Jansen, Jason D Turner, Jennifer Marshall, Moustafa Attar, Loriane Savary, Corinna Wehmeyer, Amy J Naylor, Samuel Kemble, et al. Distinct fibroblast subsets drive inflammation and damage in arthritis. *Nature*, 570(7760):246–251, 2019.

[6] Roser Vento-Tormo, Mirjana Efremova, Rachel A Botting, Margherita Y Turco, Miquel Vento-Tormo, Kerstin B Meyer, Jong-Eun Park, Emily Stephenson, Krzysztof Polański, Angela Goncalves, et al. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature*, 563(7731):347–353, 2018.

[7] Jong-Eun Park, Rachel A Botting, Cecilia Domínguez Conde, Dorin-Mirel Popescu, Marieke Lavaert, Daniel J Kunz, Issac Goh, Emily Stephenson, Roberta Ragazzini, Elizabeth Tuck, et al. A cell atlas of human thymic development defines t cell repertoire formation. *Science*, 367(6480), 2020.

[8] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381, 2014.

[9] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

[10] Zhicheng Ji and Hongkai Ji. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic acids research*, 44(13):e117–e117, 2016.

[11] Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19(1):477, 2018.

[12] Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nature methods*, 14(10):979, 2017.

[13] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC press, 1990.

[14] Simon N Wood. mgcv: Gams and generalized ridge regression for r. *R news*, 1(2):20–25, 2001.

[15] Thomas W Yee. The vgam package. *R News*, 8(2):28–39, 2008.

[16] Koen Van den Berge, Hector Roux De Bezieux, Kelly Street, Wouter Saelens, Robrecht Cannoodt, Yvan Saeys, Sandrine Dudoit, and Lieven Clement. Trajectory-based differential expression analysis for single-cell sequencing data. *Nature communications*, 11(1):1–13, 2020.

[17] Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019.

[18] Nan Hao and Erin K O'shea. Signal-dependent dynamics of transcription factor translocation controls gene expression. *Nature structural & molecular biology*, 19(1):31, 2012.

[19] Xu Ren and Pei-Fen Kuan. Negative binomial additive model for rna-seq data analysis. *BMC bioinformatics*, 21:1–15, 2020.

[20] David S Fischer, Fabian J Theis, and Nir Yosef. Impulse model-based differential expression analysis of time course sequencing data. *Nucleic acids research*, 46(20):e119–e119, 2018.

[21] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

[22] Daniel Spies, Peter F Renz, Tobias A Beyer, and Constance Ciaudo. Comparative analysis of differential gene expression tools for rna sequencing time course data. *Briefings in bioinformatics*, 20(1):288–298, 2019.

[23] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–742, 2014.

[24] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, 16(1):1–13, 2015.

[25] Trung Nghia Vu, Quin F Wills, Krishna R Kalari, Nifang Niu, Liewei Wang, Mattias Rantalainen, and Yudi Pawitan. Beta-poisson model for single-cell rna-seq data analyses. *Bioinformatics*, 32(14):2128–2135, 2016.

[26] Keegan D Korthauer, Li-Fang Chu, Michael A Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendziorski. A statistical approach for identifying differential distributions in single-cell rna-seq experiments. *Genome biology*, 17(1):222, 2016.

[27] Charlotte Soneson and Mark D Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature methods*, 15(4):255, 2018.

[28] Kieran R Campbell and Christopher Yau. Order under uncertainty: robust differential expression analysis using probabilistic models for pseudotime inference. *PLoS computational biology*, 12(11):e1005212, 2016.

[29] Magdalena E Strauß, John E Reid, and Lorenz Wernisch. Gpseudorank: a permutation sampler for single cell orderings. *Bioinformatics*, 35(4):611–618, 2019.

[30] Gregory Giecold, Eugenio Marco, Sara P Garcia, Lorenzo Trippa, and Guo-Cheng Yuan. Robust lineage reconstruction from high-dimensional single-cell data. *Nucleic acids research*, 44(14):e122–e122, 2016.

[31] Alex K Shalek, Rahul Satija, Joe Shuga, John J Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S Gertner, Jellert T Gaublomme, Nir Yosef, et al. Single-cell rna-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363–369, 2014.

[32] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[33] Wei-Lin Qiu, Yu-Wei Zhang, Ye Feng, Lin-Chen Li, Liu Yang, and Cheng-Ran Xu. Deciphering pancreatic islet β cell and α cell maturation pathways and characteristic features at the single-cell level. *Cell metabolism*, 25(5):1194–1205, 2017.

[34] Åsa Apelqvist, Hao Li, Lukas Sommer, Paul Beatus, David J Anderson, Tasuku Honjo, Martin Hrabě de Angelis, Urban Lendahl, and Helena Edlund. Notch signalling controls pancreatic cell differentiation. *Nature*, 400(6747):877–881, 1999.

[35] Rebecca Lawson, Wolfgang Maret, and Christer Hogstrand. Expression of the zip/slc39a transporters in β-cells: a systematic review and integration of multiple datasets. *BMC genomics*, 18(1):719, 2017.

[36] Franziska Paul, Ya'ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, David Lara-Astiaso, Meital Gury, Assaf Weiner, et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163(7): 1663–1677, 2015.

[37] Isaac Engel, Grégory Seumois, Lukas Chavez, Daniela Samaniego-Castruita, Brandie White, Ashu Chawla, Dennis Mock, Pandurangan Vijayanand, and Mitchell Kronenberg. Innate-like functions of natural killer t cell subsets result from highly divergent gene programs. *Nature immunology*, 17(6):728–739, 2016.

[38] Chiaowen Joyce Hsiao, PoYuan Tung, John D Blischak, Jonathan E Burnett, Kenneth A Barr, Kushal K Dey, Matthew Stephens, and Yoav Gilad. Characterizing and inferring quantitative cell cycle phase in single-cell rna-seq data analysis. *Genome research*, 30(4):611–621, 2020.

[39] Simon N Wood. On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1):221–228, 2013.

[40] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics*, 19(4):562–578, 2018.

[41] Valentine Svensson. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 38(2): 147–150, 2020.

[42] Justin D Silverman, Kimberly Roche, Sayan Mukherjee, and Lawrence A David. Naught all zeros in sequence count data are the same. *BioRxiv*, page 477794, 2020.

[43] Tapio Lönnberg, Valentine Svensson, Kylie R James, Daniel Fernandez-Ruiz, Ismail Sebina, Ruddy Montandon, Megan SF Soon, Lily G Fogg, Arya Sheela Nair, Urijah Liligeto, et al. Single-cell rna-seq and computational analysis using temporal mixture modelling resolves th1/tfh fate bifurcation in malaria. *Science immunology*, 2(9), 2017.

[44] Xiaojie Qiu, Andrew Hill, Jonathan Packer, Dejun Lin, Yi-An Ma, and Cole Trapnell. Single-cell mrna quantification and differential analysis with census. *Nature methods*, 14(3):309–315, 2017.

[45] Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, Linda Zhao, et al. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

[46] Jason D Lee, Dennis L Sun, Yuekai Sun, Jonathan E Taylor, et al. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[47] Dimitris N Politis, Joseph P Romano, and Michael Wolf. *Subsampling*. Springer Science & Business Media, 1999.

[48] George C Tseng and Wing H Wong. Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, 61(1):10–16, 2005.

[49] Yidan Sun, Heather Zhou, and Jingyi Jessica Li. Bipartite tight spectral clustering (bitsc) algorithm for identifying conserved gene co-clusters in two species. *bioRxiv*, page 865378, 2019.

[50] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[51] Simon N Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.

[52] Simon Wotherspoon and Paul Burch. *zigam: EM implementation of zero-inflated GAMs*, 2016. R package version 0.1.1.

[53] Kenneth P Burnham and David R Anderson. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

[54] Belinda Phipson and Gordon K Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1), 2010.

[55] Marie Laure Delignette-Muller and Christophe Dutang. fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4):1–34, 2015. URL http://www.jstatsoft.org/v64/i04/.

[56] Tatiana Benaglia, Didier Chauveau, David R. Hunter, and Derek Young. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29, 2009. URL http://www.jstatsoft.org/v32/i06/.

[57] Adrian Alexa and Jorg Rahnenfuhrer. *topGO: Enrichment Analysis for Gene Ontology*, 2019. R package version 2.38.1.

[58] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287, 2012. doi: 10.1089/omi.2011.0118.

[59] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications*, 9(1):1–17, 2018.

[60] Simon N. Wood. Basis dimension choice for smooths. URL `https://stat.ethz.ch/R-manual/R-patched/library/mgcv/html/choose.k.html`.

Figures

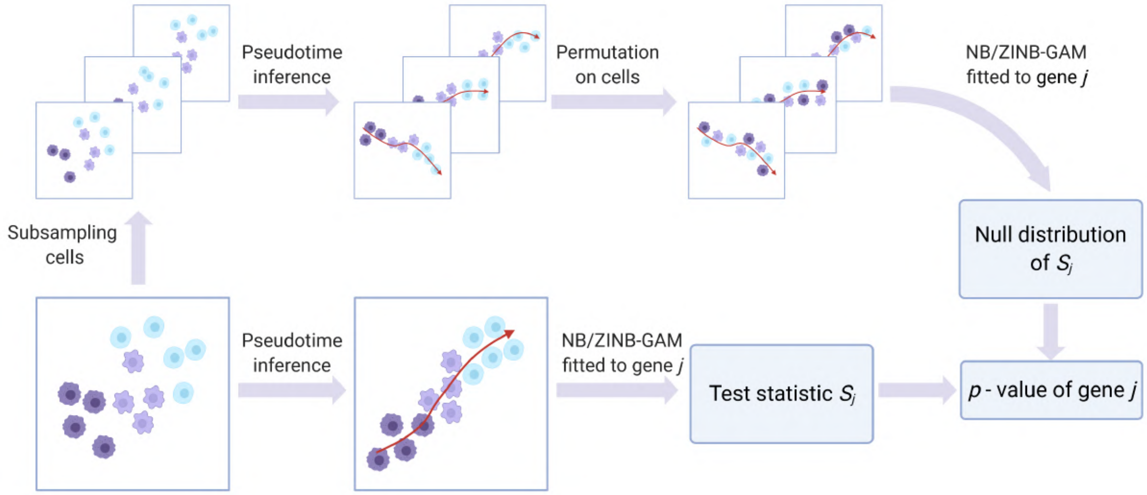


Figure 1: An illustration of the PseudotimeDE method. The core of PseudotimeDE is to obtain a valid null distribution of the DE gene test statistic S_j. To achieve that, PseudotimeDE subsamples 80% cells from the original scRNA-seq data. Then on each subsample, PseudotimeDE performs pseudotime inference (using a user-specified method such as Slingshot and Monocle3-PI) and permutes the inferred pseudotime across cells. Next, PseudotimeDE fits a model (NB-GAM or ZINB-GAM) to the permuted subsamples to obtain the values of S_j under the null hypothesis and uses these values to approximate the null distribution of S_j. In parallel, PseudotimeDE fits the same model to the original dataset and calculate the observed value of S_j. Finally, PseudotimeDE derives the p-value from the observed value and the null distribution of S_j. Detail is described in Methods.

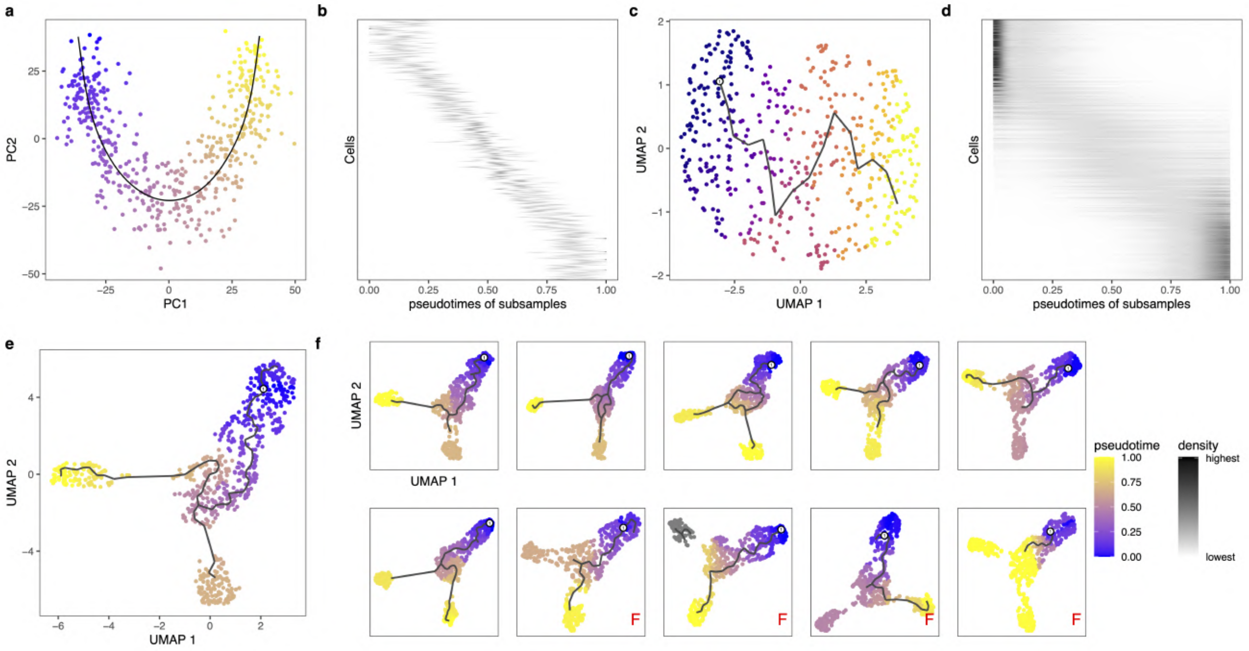


Figure 2: PseudotimeDE captures the uncertainty in pseudotime inference. (a) Visualization of synthetic single-lineage cells marked with inferred pseudotime by Slingshot (using PCA). The black curve denotes the inferred lineage. **(b)** The distributions of individual cells' inferred pseudotime by Slingshot across subsamples. In the vertical axis, cells are ordered by their true time in the lineage used in simulation; for every cell (a vertical coordinate), black dots have horizontal coordinates corresponding to the cell's inferred pseudotime in the subsamples that include the cell. The more horizontally spread out the black dots, the greater uncertainty the pseudotime inference has. **(c)** Visualization of synthetic single-lineage cells marked with inferred pseudotime by Monocle3-PI (using UMAP). The black curve denotes the inferred lineage. Compared with (a), the inferred lineage is more wiggling. **(d)** The distributions of individual cells' inferred pseudotime by Monocle3-PI across subsamples. Compared with (b), the uncertainty in pseudotime inference is greater. **(e)** Visualization of synthetic bifurcating cells marked with inferred pseudotime by Monocle3-PI (using UMAP). Monocle3-PI recovers the bifurcation topology. **(f)** Visualization of ten subsamples of the cells in (e), marked with inferred pseudotime by Monocle3-PI (using UMAP) on each subsample. Four out of the ten subsamples do not have the bifurcation topology correctly inferred (labeled with red "F"), revealing the uncertainty in pseudotime inference by Monocle3-PI. In panels (a), (c), (e) and (f), inferred pseudotime is represented by a color scale from 0 (the earliest pseudotime) to 1 (the latest pseudotime).

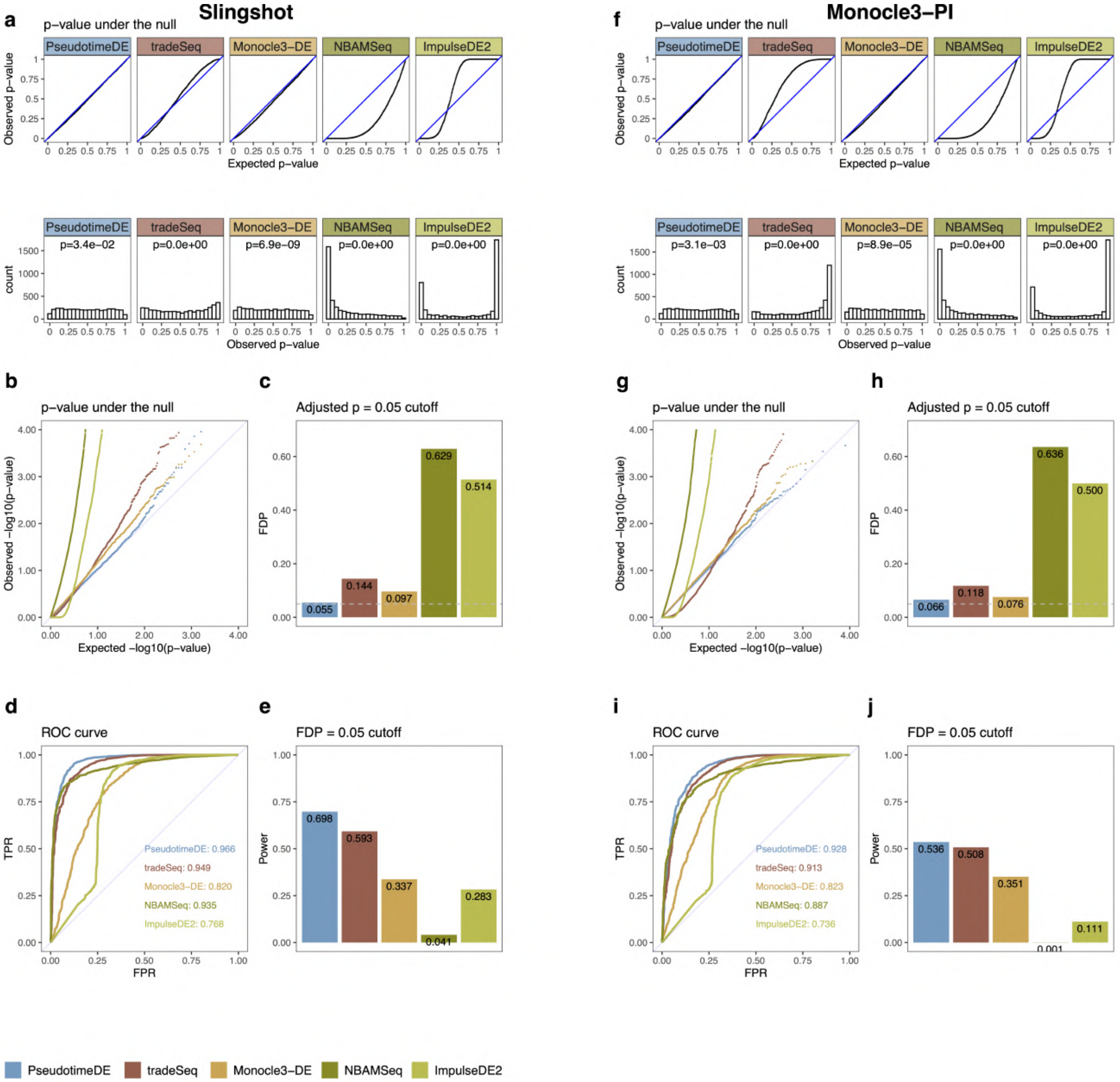
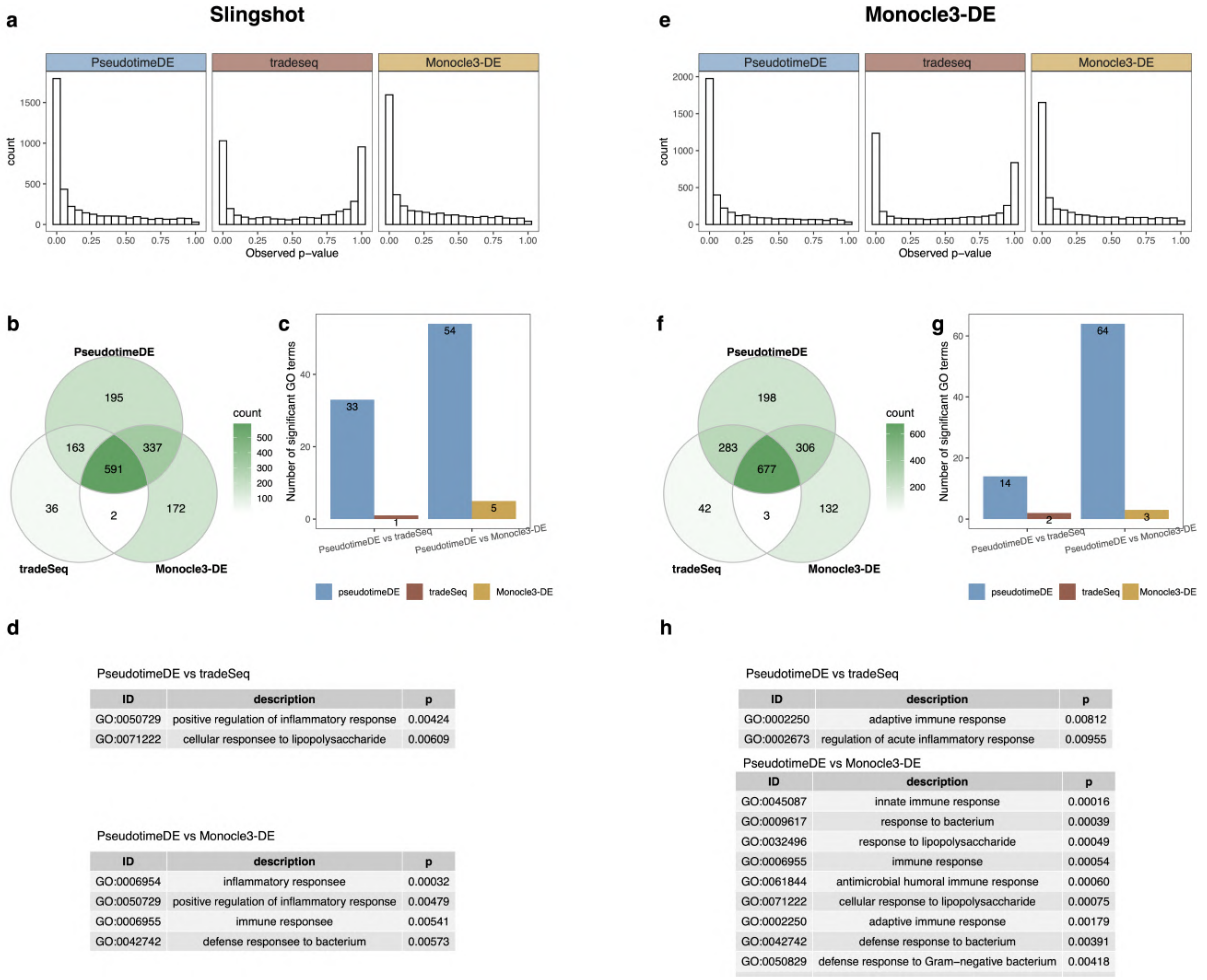


Figure 3: PseudotimeDE outperforms four state-of-the-art methods (tradeSeq, Monocle3-DE, NBAMSeq, and ImpulseDE2) for identifying DE genes along cell pseudotime. Left panels (a)–(e) are based on pseudotime inferred by Slingshot; right panels (f)–(j) are based on pseudotime inferred by Monocle3-PI. **(a) & (f)** Distributions of non-DE genes' observed p-values by five DE methods with inferred pseudotime. Top: quantile-quantile plots that compare the empirical quantiles of the observed p-values against the expected quantiles of the Uniform$[0, 1]$ distribution. Bottom: histograms of the observed p-values. The p-values shown on top of histograms are from the Kolmogorov–Smirnov test under the null hypothesis that the distribution is Uniform$[0, 1]$. The larger the p-value, the more uniform the distribution is. Among the five DE methods, PseudotimeDE's observed p-values follow most closely the expected Uniform$[0, 1]$ distribution. **(b) & (g)** Quantile-quantile plots of the same p-values as in (a) and (f) on the negative $\log_{10}$ scale. PseudotimeDE returns better-calibrated small p-values than the other four methods do. **(c) & (h)** FDPs of the five DE methods with the target FDR 0.05 (BH adjusted-$p \leq 0.05$). PseudotimeDE yields the FDP closest to 0.05. **(d) & (i)** ROC curves and AUROC values of the five DE methods. PseudotimeDE achieves the highest AUROC. **(e) & (j)** Power of the five DE methods under the FDP = 0.05 cutoff. PseudotimeDE achieves the highest power.



PseudotimeDE vs tradeSeq

ID	description	p
GO:0050729	positive regulation of inflammatory response	0.00424
GO:0071222	cellular responsee to lipopolysaccharide	0.00609

PseudotimeDE vs Monocle3-DE

ID	description	p
GO:0006954	inflammatory responsee	0.00032
GO:0050729	positive regulation of inflammatory response	0.00479
GO:0006955	immune responsee	0.00541
GO:0042742	defense responsee to bacterium	0.00573

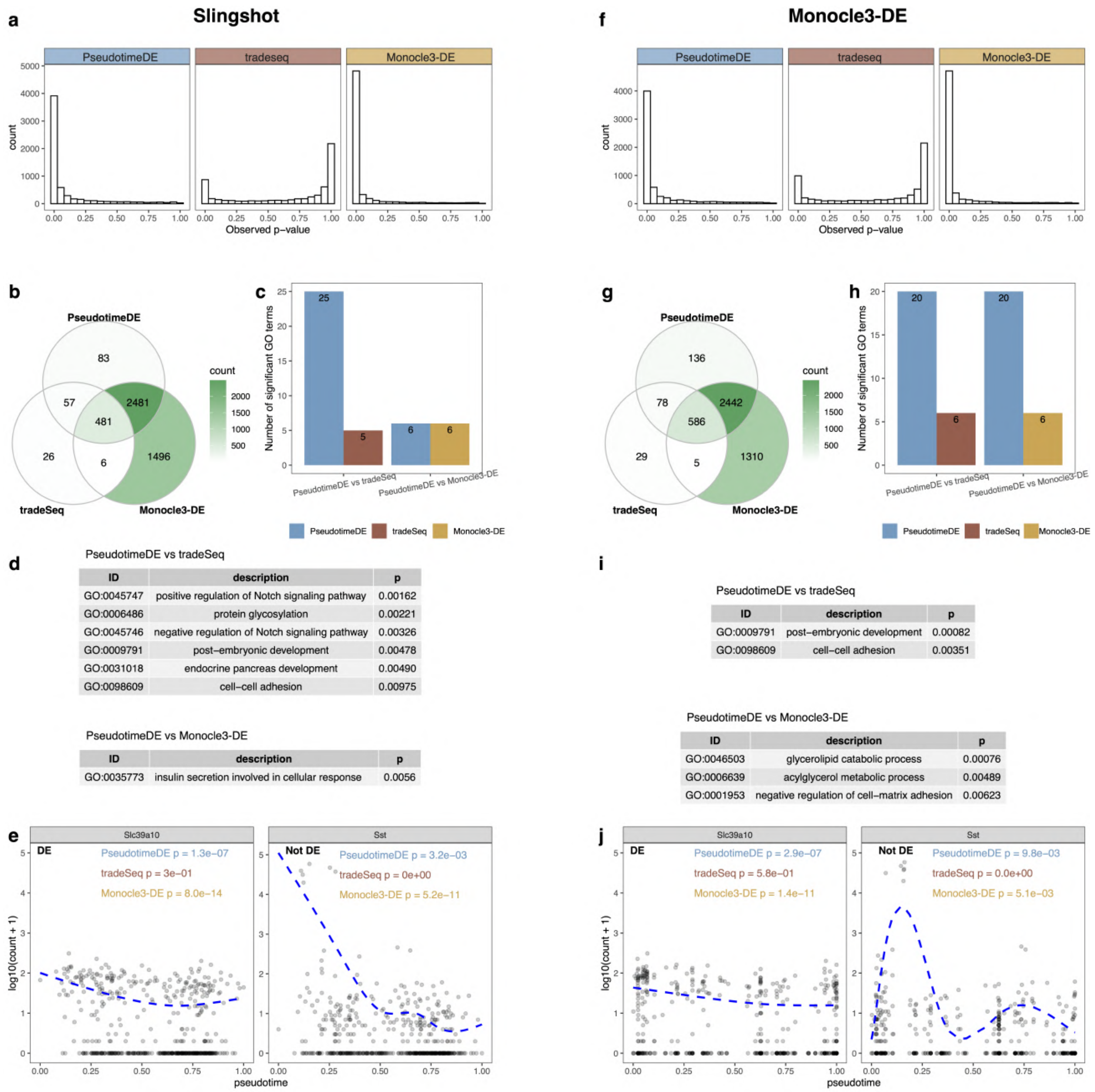
PseudotimeDE vs tradeSeq

ID	description	p
GO:0002250	adaptive immune response	0.00812
GO:0002673	regulation of acute inflammatory response	0.00955

PseudotimeDE vs Monocle3-DE

ID	description	p
GO:0045087	innate immune response	0.00016
GO:0009617	response to bacterium	0.00039
GO:0032496	response to lipopolysaccharide	0.00049
GO:0006955	immune response	0.00054
GO:0061844	antimicrobial humoral immune response	0.00060
GO:0071222	cellular response to lipopolysaccharide	0.00075
GO:0002250	adaptive immune response	0.00179
GO:0042742	defense response to bacterium	0.00391
GO:0050829	defense response to Gram–negative bacterium	0.00418

Figure 4: Application of PseudotimeDE, tradeSeq, and Monocle3-DE to the LPS-dendritic cell dataset. Left panels (a)–(d) are based on pseudotime inferred by Slingshot; right panels (e)–(h) are based on pseudotime inferred by Monocle3-PI. **(a) & (e)** Histograms of all genes' p-values by the three DE methods. The bimodal distributions of tradeSeq's p-values suggest a violation of the requirement that p-values follow the Uniform$[0, 1]$ distribution under the null hypothesis. **(b) & (f)** Venn plots showing the overlaps of the significant DE genes (BH adjusted-$p \leq 0.01$) identified by the three DE methods. PseudotimeDE's DE genes nearly include tradeSeq's. **(c) & (g)** Numbers of GO terms enriched $(p < 0.01)$ in the significant DE genes specifically found by PseudotimeDE or tradeSeq/Monocle3-DE in pairwise comparisons between PseudotimeDE and tradeSeq/Monocle3-DE in (b) & (f). Many more GO terms are enriched in the PseudotimeDE-specific DE genes than in the tradeSeq- or Monocle3-DE-specific ones. **(d) & (h)** Example GO terms enriched in the Pseudotime-specific DE genes in (c) & (g). Many of these terms are related to LPS, immune process, and defense to bacterium.



(d) PseudotimeDE vs tradeSeq

ID	description	p
GO:0045747	positive regulation of Notch signaling pathway	0.00162
GO:0006486	protein glycosylation	0.00221
GO:0045746	negative regulation of Notch signaling pathway	0.00326
GO:0009791	post−embryonic development	0.00478
GO:0031018	endocrine pancreas development	0.00490
GO:0098609	cell−cell adhesion	0.00975

PseudotimeDE vs Monocle3-DE

ID	description	p
GO:0035773	insulin secretion involved in cellular response	0.0056

(i) PseudotimeDE vs tradeSeq

ID	description	p
GO:0009791	post−embryonic development	0.00082
GO:0098609	cell−cell adhesion	0.00351

PseudotimeDE vs Monocle3-DE

ID	description	p
GO:0046503	glycerolipid catabolic process	0.00076
GO:0006639	acylglycerol metabolic process	0.00489
GO:0001953	negative regulation of cell−matrix adhesion	0.00623

Figure 5: Application of PseudotimeDE, tradeSeq, and Monocle3-DE to the pancreatic beta cell maturation dataset. Left panels (a)–(e) are based on pseudotime inferred by Slingshot; right panels (f)–(j) are based on pseudotime inferred by Monocle3-PI. **(a) & (f)** Histograms of all genes' p-values by the three DE methods. The bimodal distributions of tradeSeq's p-values suggest a violation of the requirement that p-values follow the Uniform$[0, 1]$ distribution under the null hypothesis. **(b) & (g)** Venn plots showing the overlaps of the significant DE genes (BH adjusted-$p \leq 0.01$) identified by the three DE methods. PseudotimeDE's DE genes nearly include tradeSeq's. **(c) & (h)** Numbers of GO terms enriched ($p < 0.01$) in the significant DE genes specifically found by PseudotimeDE or tradeSeq/Monocle3-DE in pairwise comparisons between PseudotimeDE and tradeSeq/Monocle3-DE in (b) & (g). Many more GO terms are enriched in the PseudotimeDE-specific DE genes than in the tradeSeq- or Monocle3-DE-specific ones. **(d) & (i)** Example GO terms enriched in the Pseudotime-specific DE genes in (c) & (h). Many of these terms are related to related to insulin, beta cell regulation, and pancreas development. **(e) & (j)** Two examples genes: *Slc39a10* (DE) and *Sst* (non-DE). For *Slc39a10*, both PseudotimeDE and Monocle3-DE yield small p-values ($p < 1e-6$), while tradeSeq does not ($p > 0.1$). For *Sst*, PseudotimeDE yields larger p-values than tradeSeq and Monocle3-DE do. Dashed blue lines are the fitted curves by NB-GAM.

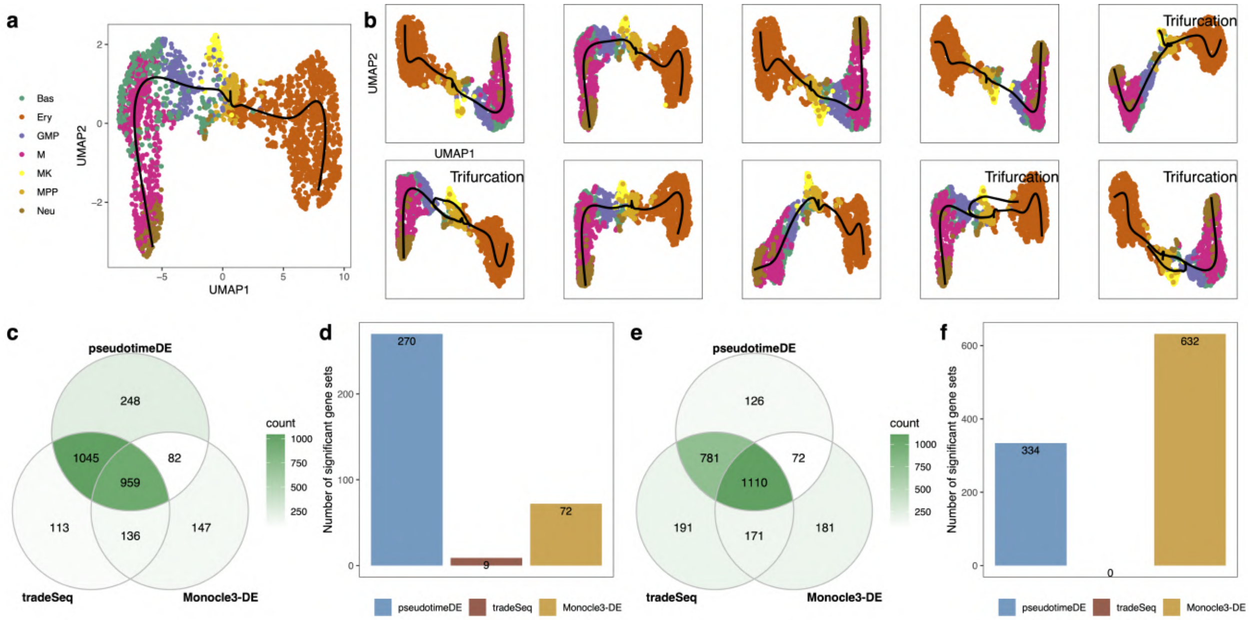


Figure 6: Application of PseudotimeDE, tradeSeq, and Monocle3-DE to the mouse bone marrow dataset. (a) UMAP visualization and inferred pseudotime by Slingshot. Pre-defined cell types are marked by colors. Slingshot returns a bifurcation topology, denoted as lineage 1 (left) and lineage 2 (right). **(b)** UMAP visualization and inferred pseudotime by Slingshot on ten random subsamples. Four out of ten subsamples do not yield bifurcation topology but trifurcation topology, where the third lineage mainly contains the cell type "MK" and was reported in [12]. **(c)** Histograms of all genes' p-values calculated by the three DE methods in the first lineage. **(d)** Histograms of all genes' p-values calculated by the three DE methods in the second lineage. **(e)** Venn plot showing the overlaps of the significant DE genes (BH adjusted-$p \leq 0.01$) identified by the three DE methods in lineage 1. PseudotimeDE and tradeSeq share 77.6% (Jaccard index) DE genes. **(f)** Numbers of enriched gene sets ($q < 0.25$) by GSEA using the p-values in lineage 1 by the three DE methods. Although the DE genes are similar in (e), PseudotimeDE yields 270 enriched gene sets, while tradeSeq only yields 9. **(g)** Venn plot showing the overlaps of the significant DE genes (BH adjusted-$p \leq 0.01$) identified by the three DE methods in lineage 2. Similar to lineage 1 in (g), PseudotimeDE and tradeSeq share 77.2% (Jaccard index) DE genes. **(h)** Numbers of enriched gene sets ($q < 0.25$) by GSEA using the p-values in lineage 2 by the three DE methods. PseudotimeDE and Monocle3-DE yield hundreds of enriched gene sets, while tradeSeq does not yield any enriched gene sets.

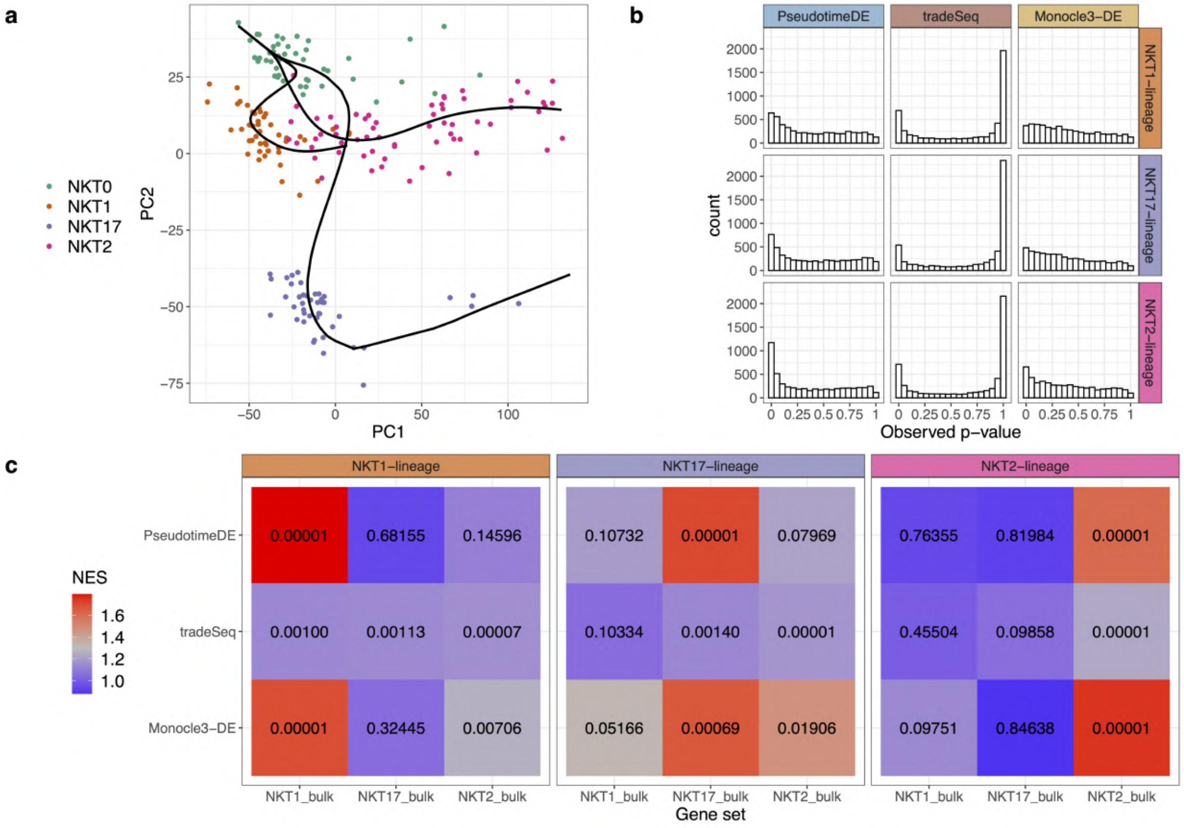


Figure 7: Application of PseudotimeDE, tradeSeq, and Monocle3-DE to the natural killer T cell dataset. (a) PCA visualization and inferred pseudotime by Slingshot. Pre-defined NKT subtypes are marked by colors. Slingshot returns a trifurcation topology, where the three lineages are NKT0 to NKT1, NKT0 to NKT17, and NKT0 to NKT2. **(b)** Histograms of all genes' p-values in the three lineages calculated by the three DE methods. **(c)** Heatmaps of normalized enrichment scores (NESs, marked by colors) and their corresponding p-values (in numbers) from the GSEA analysis. Each NES value and its corresponding p-value are calculated for each DE method and each lineage, based on the p-values of a DE method for a lineage and that lineage's DE genes found from bulk RNA-seq data, denoted by "NKT1 bulk", "NKT17 bulk," or "NKT2 bulk" [37]. Note that among the three DE methods, PseudotimeDE outputs p-values that best agree with the lineage-specific DE genes from bulk data and thus most distinguish the three lineages. For instance, for the NKT1 lineage, PseudotimeDE's small p-values are enriched in the "NKT1 bulk" gene set only, while tradeSeq and Monocle3-DE have small p-values enriched in at least two lineage-specific DE gene sets.

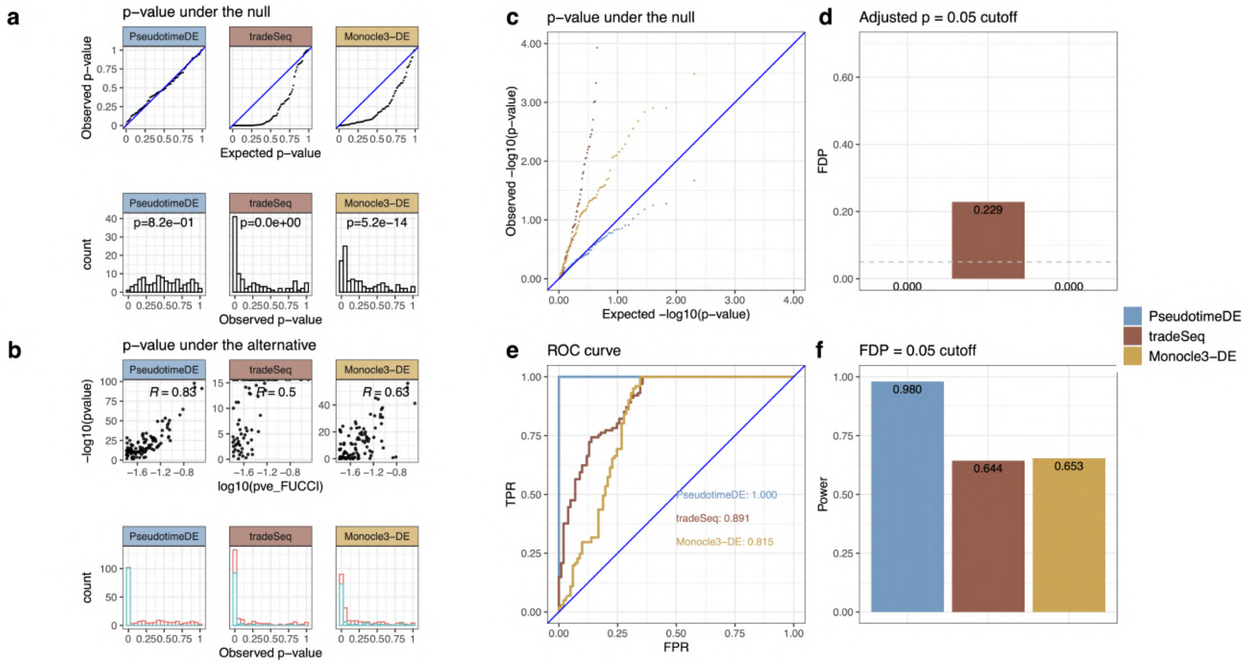


Figure 8: Application of PseudotimeDE, tradeSeq, and Monocle3-DE to the cell cycle phase dataset. (a) Distributions of non-DE genes' p-values by three DE methods with inferred pseudotime. Top: quantile-quantile plots that compare the empirical quantiles of non-DE genes' p-values against the expected quantiles of the Uniform$[0,1]$ distribution. Bottom: histograms of non-DE genes' p-values. The p-values shown on top of histograms are from the Kolmogorov–Smirnov test under the null hypothesis that the distribution is Uniform$[0,1]$. The larger the p-value, the more uniform the distribution is. Among the three DE methods, PseudotimeDE's p-values follow most closely the expected Uniform$[0,1]$ distribution. **(b)** Distributions of DE genes' p-values by three DE methods with inferred pseudotime. Top: scatter plots of DE genes' p-values against the proportions of variance explained (PVE), which measure the strengths of genes' inferred cyclic trends in the original study [38]. PseudotimeDE's p-values ($-\log_{10}$ transformed) have the highest correlation with the PVE, indicating that PseudotimeDE identifies the genes with the strongest cyclic trends as the top DE genes. Bottom: histograms of all genes' p-values. Blue and red colors represent the p-values of DE genes and non-DE genes (same as in (a) bottom), respectively. PseudotimeDE yields the best separation of the two gene groups' p-values. **(c)** Quantile-quantile plots of the same p-values as in (a) on the negative $\log_{10}$ scale. PseudotimeDE returns the best-calibrated p-values. **(d)** FDPs of the three DE methods with the target FDR 0.05 (BH adjusted-$p \leq 0.05$). **(e)** ROC curves and AUROC values of the three DE methods. PseudotimeDE achieves the highest AUROC. **(f)** Power of the three DE methods under the FDP = 0.05 cutoff. PseudotimeDE achieves the highest power.

Supplementary Materials

Goodness-of-fit of the parametric distribution

We perform a goodness-of-fit test to show that our parametric approach fits the empirical distributions well (Fig. S21). We choose the Anderson-Darling (AD) test since it gives more weights on the tail of the distribution than the Kolmogorov-Smirnov test does. On one simulated dataset and one real dataset, the distribution of AD test p-values is approximately Uniform$[0, 1]$ (Fig. S21a & c), indicating that the parametric distributions fit the empirical null distributions well (i.e., all cases are from the null). Since this goodness-of-fit check is crucial, `PseudotimeDE` R package automatically outputs the AD test p-value per gene so that users can check whether the parametric approximation works well; if not, users may increase the number of subsampling and use the empirical null for DE p-value calculation.

Choice of the subsampling proportion

In general, results of PseudotimeDE is robust to the subsampling proportion. We have applied to one simulation dataset (Fig. S22) five sampling proportions (50%, 60%, 70%, 80%, and 90%), among which 80% is the default used in PseudotimeDE. From the results, we first observe that the p-values (on the $-\log_{10}$ scale) are highly correlated between 80% and any other proportion (Fig. S22c), confirming the robustness of PseudotimeDE to the subsampling proportion. Second, we observe that the power of PseudotimeDE remains stable under varying sampling proportions, though the power is slightly lower when the sampling proportion is 50% (Fig. S22a). Third, the false discovery proportion (FDP) of PseudotimeDE is also stable under varying sampling proportions, though the FDP is slightly higher than the nominal level when the sampling proportion is 90% (Fig. S22b). Therefore, we conclude that 80% is a reasonable default sampling proportion.

Differences of the NB-GAM fitting in PseudotimeDE, tradeSeq, and NBAMSeq

PseudotimeDE, tradeSeq, and NBAMSeq all rely on the R package `mgcv` for fitting NB-GAM by penalized-restricted maximum likelihood. Therefore, the model fitting of NB-GAM is overall similar in the three methods. There are only minor differences, listed below.

1. In PseudotimeDE, the dispersion parameter ϕ_j of gene j is estimated simultaneously with the mean parameter μ_{ij} of gene j in cell i by the penalized-restricted maximum likelihood (PRML); in contrast, NBAMSeq uses the maximum a posterior (MAP) dispersion estimate by R package `DESeq2` [21]. NBAMSeq is designed for bulk RNA-seq and accounts for a samll sample size (i.e., number of replicates); therefore, the MAP estimate by `DESeq2` is more biased (for variance reduction purpose) than the PRML estimate used by PseudotimeDE.

Given that scRNA-seq has a large enough sample size (i.e., number of cells), we think that the PRML estimate of ϕ_j is preferred over the MAP estimate.

2. In PseudotimeDE, the number of knots K is pre-defined as a fix number (the default $K = 6$, which is usually large enough); In contrast, tradeSeq randomly subsamples a small set of genes to choose a K for all genes.

3. PseudotimeDE fits a NB-GAM to each lineage, while tradeSeq incorporates a lineage covariate so that it can fit a NB-GAM to multiple lineages jointly. The reason is that PseudotimeDE needs to account for the uncertainty in pseudotime inference and within-lineage model fitting makes the inference easier.

4. Pseudotime also implements a zero-inflated NB-GAM (ZINB-GAM) that includes a dropout probability parameter for each gene, and PseudotimeDE estimates this parameter internally. In contrast, while tradeSeq can also turn its NB-GAM into ZINB-GAM, it estimates the dropout probability parameters externally by ZINB-WaVE [59]; NBAMSeq is designed for bulk RNA-seq data and does not consider zero inflation.

Although the NB-GAM fitting of PseudotimeDE is similar to that of the other two methods, its inference (p-value calculation) is totally different, and this is the main novelty of PseudotimeDE.

The number of knots K

The number of knots, K, a parameter in the generalized additive model (GAM), defines the number and positions of piecewise polynomial functions in a spline function.

$$f_j(T_i) = \sum_{k=1}^{K} b_k(T_i)\beta_{jk}$$

where $f_j(\cdot)$ is the smooth spline function of gene j, T_i is the pseudotime of cell i, and K is the number of knots. It is worth noting that $K-1$ is also the **nominal degree of freedom** of the model.

K needs to be specified before the negative binomial-GAM (NB-GAM) is fitted. For the GAM model, "exact choice of K is not generally critical [60]," if the GAM model is correctly used for inference. This is why we do not include the K selection step in PseudotimeDE. The default $K = 6$ in PseudotimeDE is usually large enough for modelling gene trajectories. Using a larger K is not harmful, but it will increase the computational time.

In contrast, tradeSeq has a K selection step, which we think is problematic. To choose K, tradeSeq randomly samples a small set of genes and uses AIC to find an "appropriate" K for all genes (see *Choosing an appropriate number of knots* in the tradeSeq paper [16]). We follow their vignette and obtain $K = 4$ in Fig. 3a and $K = 6$ in Fig. 3f. In tradeSeq, the null distribution of the test statistic S_j is

$$S_j \sim \chi^2_{K-1}$$

Therefore, when $K = 4$ (Fig. 3a), the null distribution of every gene is χ^2_3; when $K = 6$ (Fig. 3f), the null distribution of every gene is χ^2_5. As a result, the distributions of tradeSeq p-values under the null are quite different between Fig. 3a and Fig. 3f. It seems like the results of tradeSeq is highly sensitive to the choice of K. Such sensitivity is probably due to their incorrect null distribution. That is, tradeSeq uses the nominal degree of freedom $K - 1$ in their null distribution (χ^2_{K-1}); however, the correct way is to use the effective degree of freedom, which is gene-specific and always smaller than $K - 1$ due to the penalization on the likelihood function. As a result, tradeSeq uses null distributions (of its test statistics) that have heavier right tails than those of the correct null distributions, resulting in conservative p-values (whose distribution under the null has a mode near 1). Details about the theory and usage of the effective degree of freedom in GAM can be found in Dr. Simon Woods' book [51] and 2013 *Biometrika* paper [39].

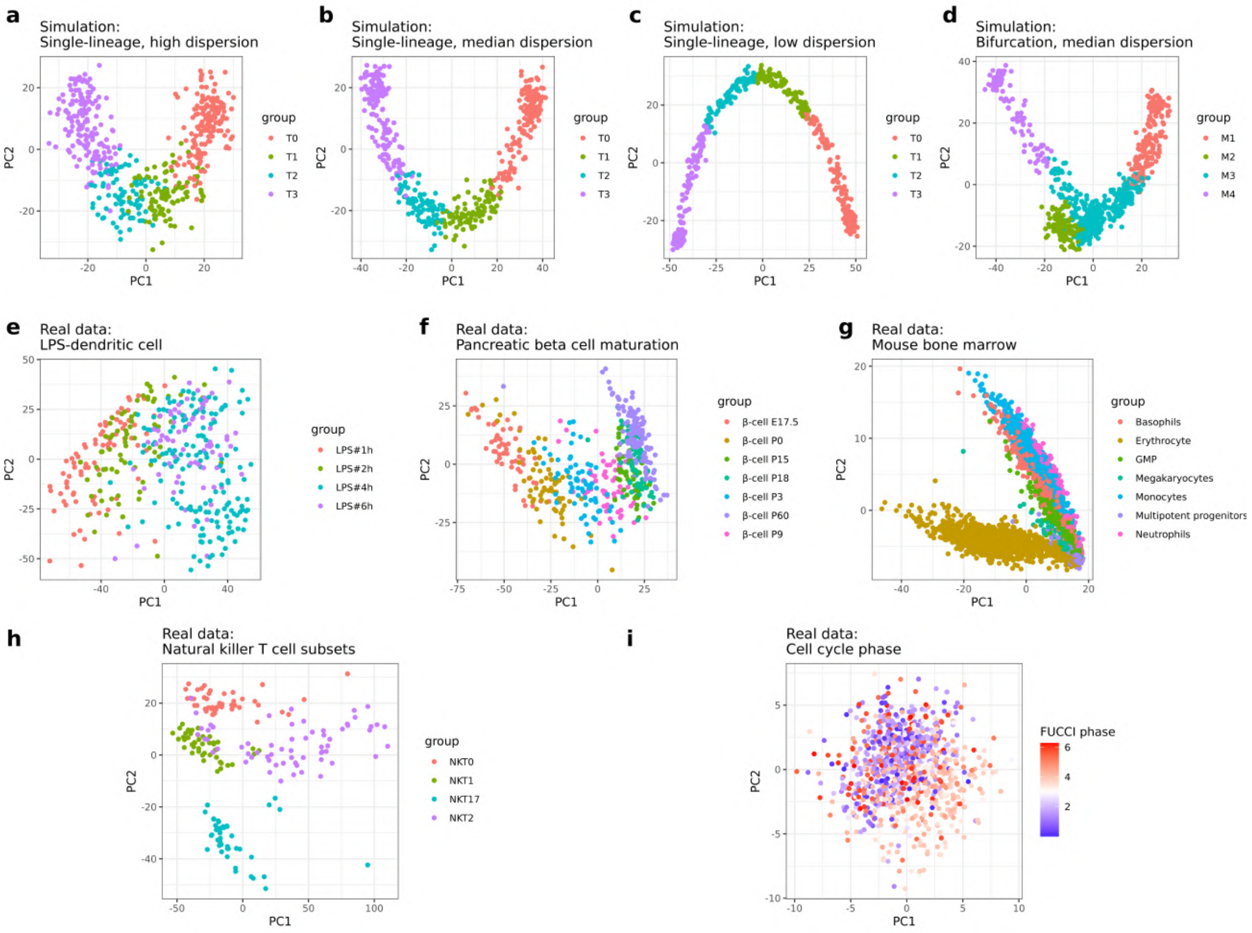


Figure S1: PCA visualization of datasets. PCA visualization of all synthetic and real datasets used in this paper. Panels (a)–(d) are synthetic datasets, where the groups correspond to time points. Panels (e)–(h) are real datasets, where the groups correspond to time points in (e) & (f) or annotated cell types in (g) & (h). Panel (i) is a real dataset with external cell cycle information, where the FUCCI phase indicates experimentally measured cell cycle phase.

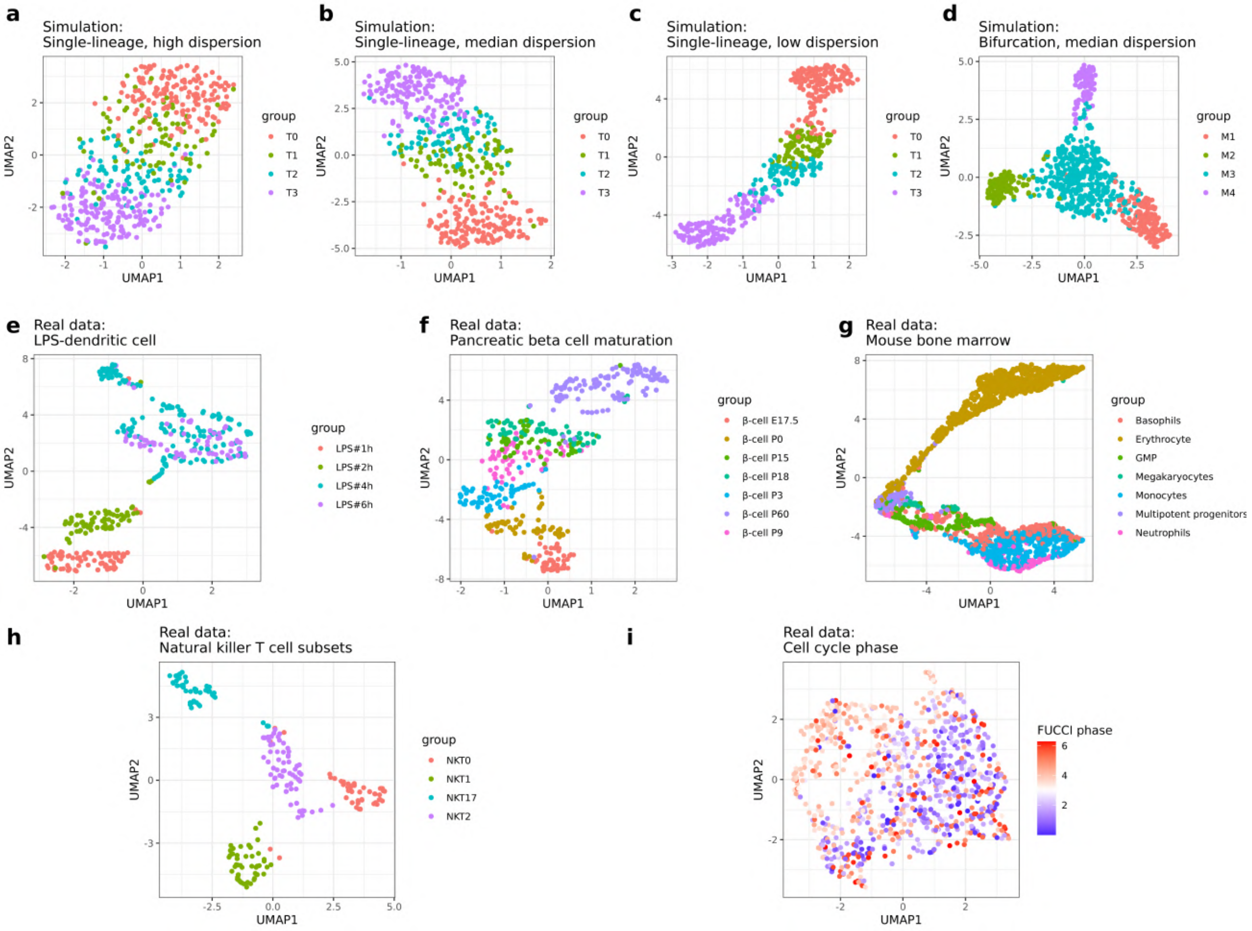


Figure S2: UMAP visualization of datasets. UMAP visualization of all synthetic and real datasets used in this paper. Panels (a)–(d) are synthetic datasets, where the groups correspond to time points. Panels (e)–(h) are real datasets, where the groups correspond to time points in (e) & (f) or annotated cell types in (g) & (h). Panel (i) is a real dataset with external cell cycle information, where the FUCCI phase indicates experimentally measured cell cycle phase.

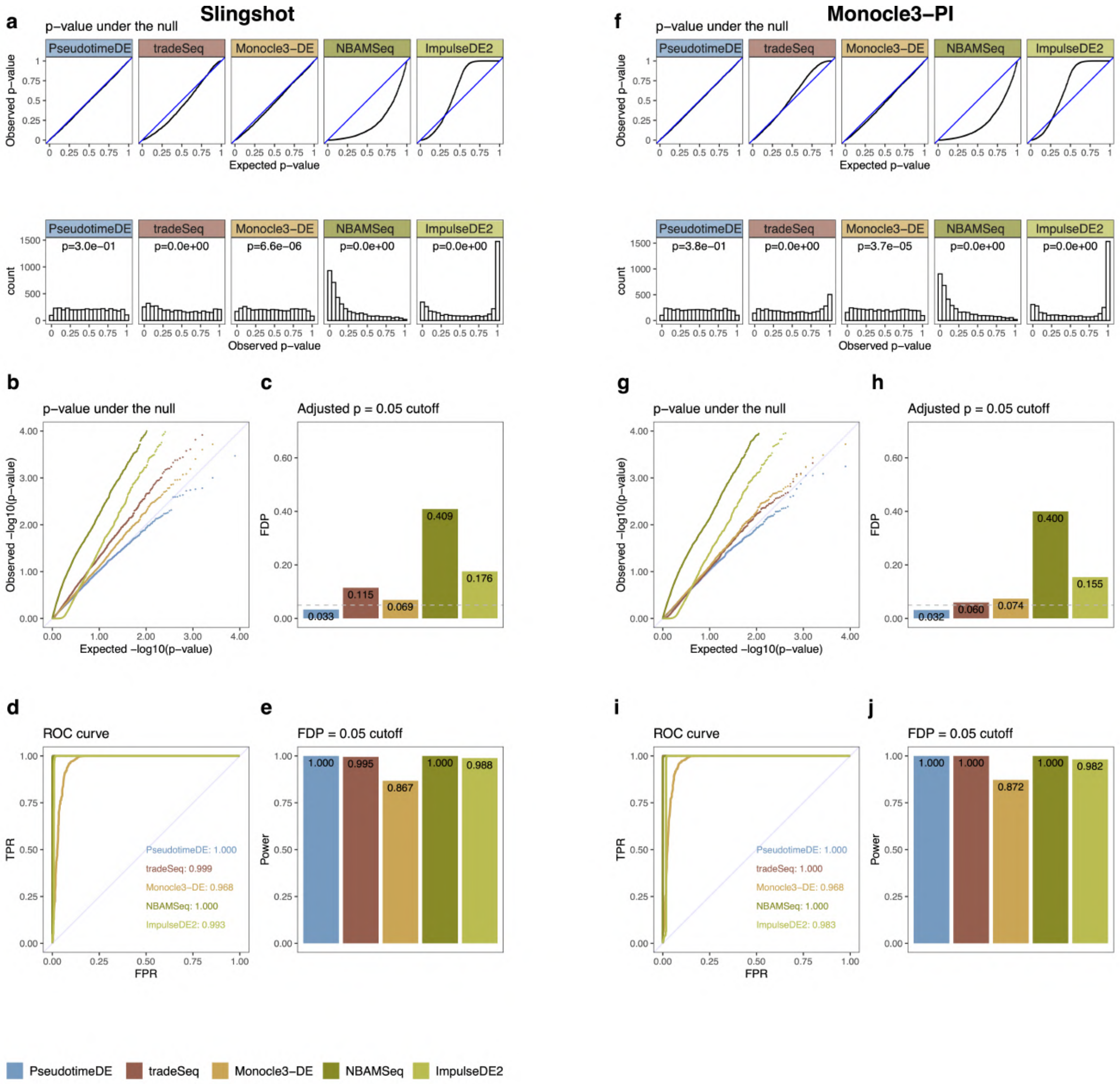


Figure S3: Comparison of five methods (PseudotimeDE, tradeSeq, Monocle3-DE, NBAMSeq, ImpulseDE2) for identifying DE genes along cell pseudotime on synthetic single-lineage data with low dispersion. Left panels (a)–(e) are based on pseudotime inferred by Slingshot; right panels (f)–(j) are based on pseudotime inferred by Monocle3-PI. **(a) & (f)** Distributions of non-DE genes' observed p-values by five DE methods with inferred pseudotime. Top: quantile-quantile plots that compare the empirical quantiles of the observed p-values against the expected quantiles of the Uniform$[0, 1]$ distribution. Bottom: histograms of the observed p-values. The p-values shown on top of histograms are from the Kolmogorov–Smirnov test under the null hypothesis that the distribution is Uniform$[0, 1]$. The larger the p-value, the more uniform the distribution is. Among the five DE methods, PseudotimeDE's observed p-values follow most closely the expected Uniform$[0, 1]$ distribution. **(b) & (g)** Quantile-quantile plots of the same p-values as in (a) and (f) on the negative $\log_{10}$ scale. PseudotimeDE returns better-calibrated small p-values than the other four methods do. **(c) & (h)** FDPs of the five DE methods with the target FDR 0.05 (BH adjusted-$p \leq 0.05$). PseudotimeDE yields the FDP below 0.05 while other methods do not. **(d) & (i)** ROC curves and AUROC values of the five DE methods. Since the dispersion of data is extremely low, all methods achieves high AUROC. **(e) & (j)** Power of the five DE methods under the FDP = 0.05 cutoff. Due to the same low dispersion reason, all methods achieves high power.

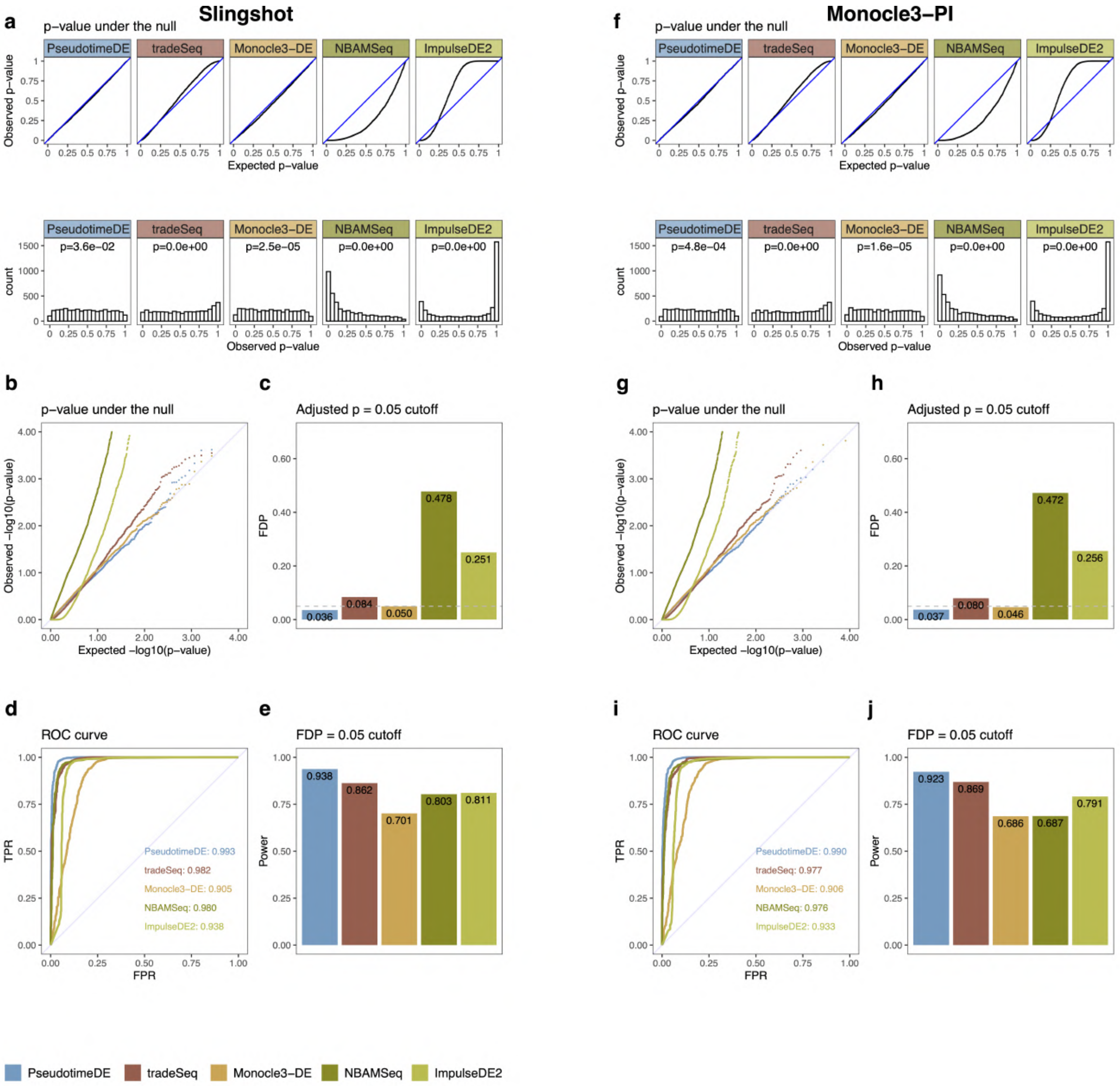


Figure S4: Comparison of five methods (PseudotimeDE, tradeSeq, Monocle3-DE, NBAMSeq, ImpulseDE2) for identifying DE genes along cell pseudotime on synthetic single-lineage data with median dispersion. Left panels (a)–(e) are based on pseudotime inferred by Slingshot; right panels (f)–(j) are based on pseudotime inferred by Monocle3-PI. **(a) & (f)** Distributions of non-DE genes' observed p-values by five DE methods with inferred pseudotime. Top: quantile-quantile plots that compare the empirical quantiles of the observed p-values against the expected quantiles of the Uniform$[0, 1]$ distribution. Bottom: histograms of the observed p-values. The p-values shown on top of histograms are from the Kolmogorov–Smirnov test under the null hypothesis that the distribution is Uniform$[0, 1]$. The larger the p-value, the more uniform the distribution is. Among the five DE methods, PseudotimeDE's observed p-values follow most closely the expected Uniform$[0, 1]$ distribution. **(b) & (g)** Quantile-quantile plots of the same p-values as in (a) and (f) on the negative $\log_{10}$ scale. PseudotimeDE returns better-calibrated small p-values than the other four methods do. **(c) & (h)** FDPs of the five DE methods with the target FDR 0.05 (BH adjusted-$p \leq 0.05$).PseudotimeDE yields the FDP below 0.05 while other methods do not. **(d) & (i)** ROC curves and AUROC values of the five DE methods. PseudotimeDE achieves the highest AUROC. **(e) & (j)** Power of the five DE methods under the FDP = 0.05 cutoff. PseudotimeDE achieves the highest power.

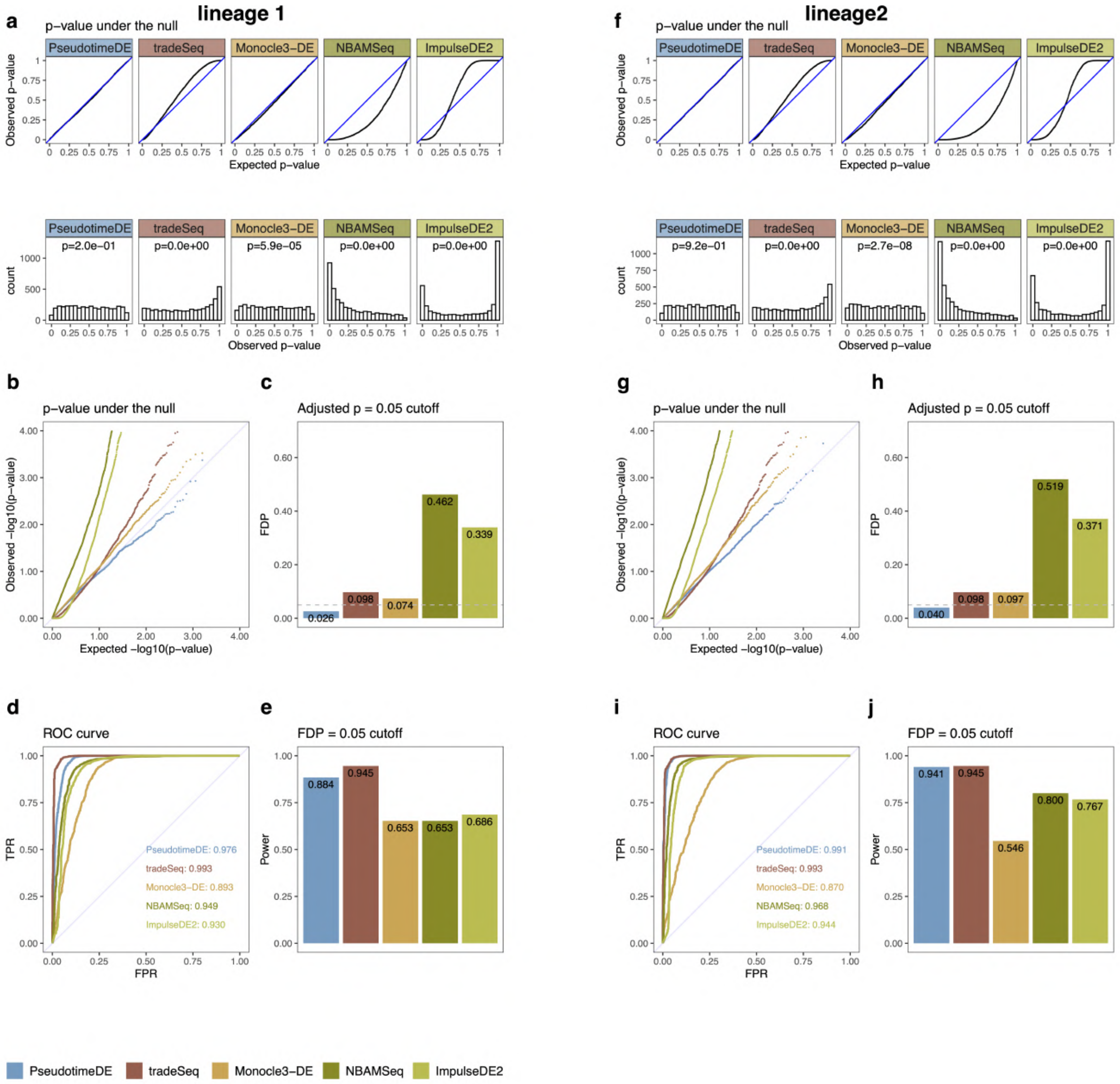


Figure S5: Comparison of five methods (PseudotimeDE, tradeSeq, Monocle3-DE, NBAMSeq, ImpulseDE2) for identifying DE genes along cell pseudotime on synthetic bifurcation data. Pseudotime is inferred by Slingshot. Left panels (a)–(e) are based on lineage 1 of two lineages in bifurcation data; right panels (f)–(j) are based on lineage 2. **(a) & (f)** Distributions of non-DE genes' observed p-values by five DE methods with inferred pseudotime. Top: quantile-quantile plots that compare the empirical quantiles of the observed p-values against the expected quantiles of the Uniform$[0, 1]$ distribution. Bottom: histograms of the observed p-values. The p-values shown on top of histograms are from the Kolmogorov–Smirnov test under the null hypothesis that the distribution is Uniform$[0, 1]$. The larger the p-value, the more uniform the distribution is. Among the five DE methods, PseudotimeDE's observed p-values follow most closely the expected Uniform$[0, 1]$ distribution. **(b) & (g)** Quantile-quantile plots of the same p-values as in (a) and (f) on the negative $\log_{10}$ scale. PseudotimeDE returns better-calibrated small p-values than the other four methods do. **(c) & (h)** FDPs of the five DE methods with the target FDR 0.05 (BH adjusted-$p \leq 0.05$). PseudotimeDE yields the FDP below 0.05 while other methods do not. **(d) & (i)** ROC curves and AUROC values of the five DE methods. PseudotimeDE achieves the second highest AUROC which is close to the highest value. **(e) & (j)** Power of the five DE methods under the FDP = 0.05 cutoff. PseudotimeDE achieves the second highest power which is close to the highest value.

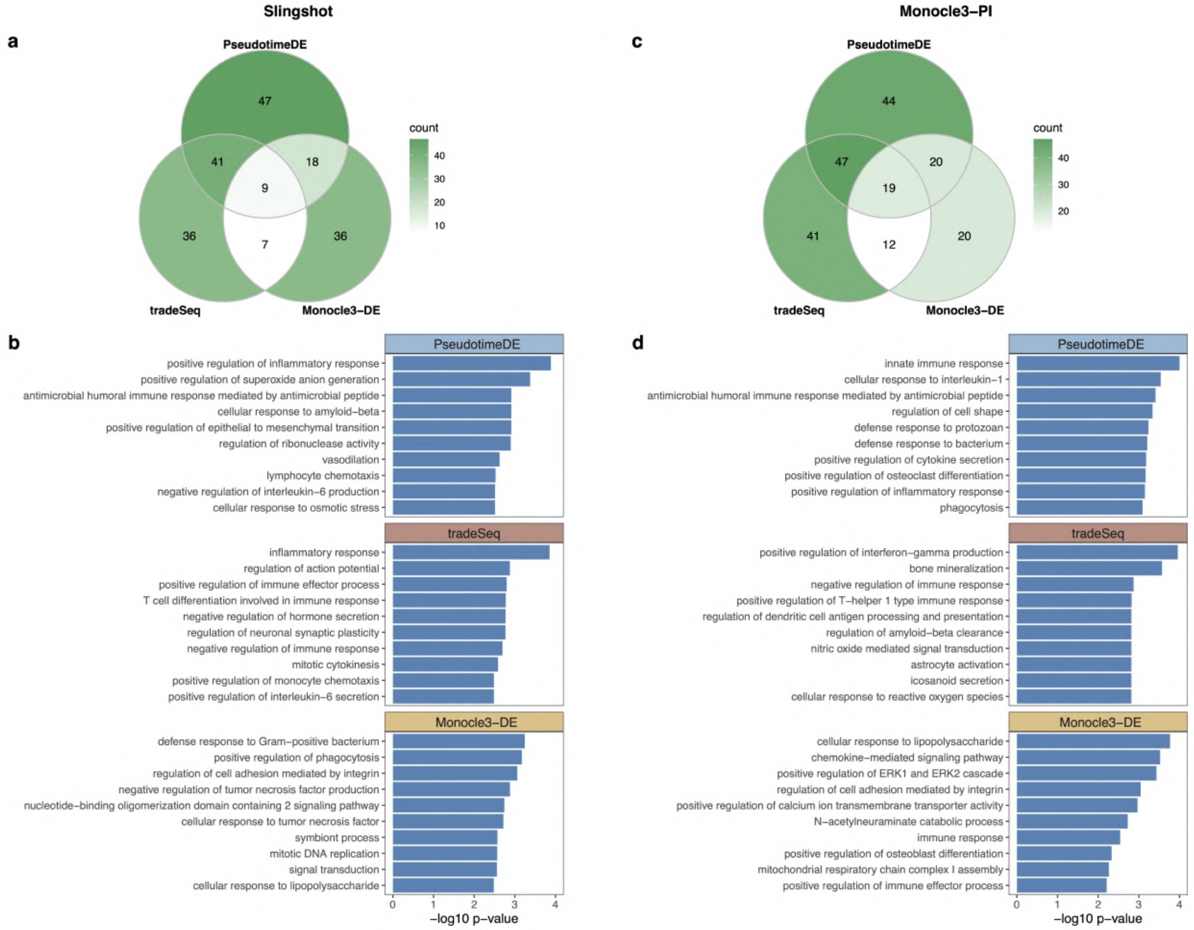
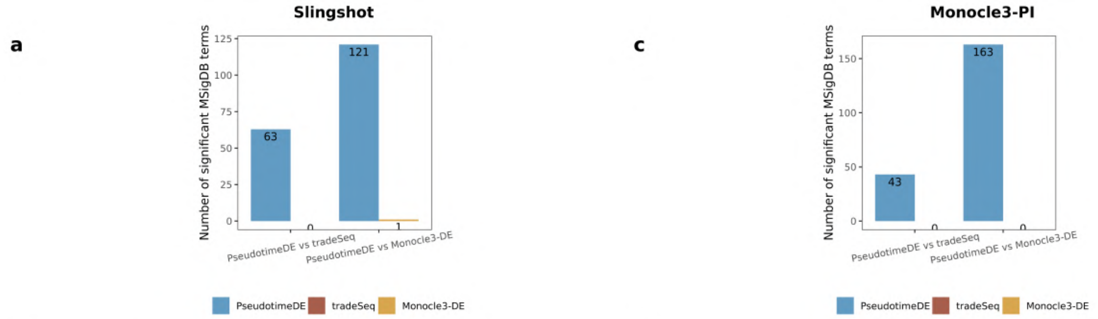


Figure S6: GO analysis of DE genes identified in the LPS-dendritic cell dataset. Left panels (a)–(b) are based on pseudotime inferred by Slingshot; right panels (c)–(d) are based on pseudotime inferred by Monocle3-PI. **(a) & (c)** Numbers of GO terms enriched ($p < 0.01$) in the significant DE genes found by each method in Fig. 4 (b) & (f). **(b) & (d)** Top 10 enriched GO terms for each DE method.



b

PseudotimeDE vs tradeSeq

description	p	p.adj
GSE2706_UNSTIM_VS_2H_LPS_AND_R848_DC_DN	1.2e-05	0.0024
GSE2706_2H_VS_8H_R848_AND_LPS_STIM_DC_DN	2.6e-05	0.0036
GSE14000_4H_VS_16H_LPS_DC_TRANSLATED_RNA_UP	6.4e-05	0.0062

PseudotimeDE vs Monocle3-DE

description	p	p.adj
GSE2706_UNSTIM_VS_2H_LPS_AND_R848_DC_DN	4.1e-07	0.00011
GSE2706_R848_VS_R848_AND_LPS_2H_STIM_DC_DN	3.0e-06	0.00040
GSE14000_UNSTIM_VS_4H_LPS_DC_DN	4.2e-06	0.00051
GSE14000_UNSTIM_VS_4H_LPS_DC_TRANSLATED_RNA_DN	7.1e-06	0.00072
GSE2706_R848_VS_LPS_2H_STIM_DC_DN	5.6e-05	0.00338

d

PseudotimeDE vs tradeSeq

description	p	p.adj
GSE14000_UNSTIM_VS_16H_LPS_DC_TRANSLATED_RNA_UP	4.8e-05	0.0058

PseudotimeDE vs Monocle3-DE

description	p	p.adj
GSE2706_UNSTIM_VS_2H_LPS_DC_DN	7.8e-08	3.1e-05
GSE14000_UNSTIM_VS_4H_LPS_DC_DN	9.9e-07	1.6e-04
GSE2706_UNSTIM_VS_2H_LPS_AND_R848_DC_DN	9.9e-07	1.6e-04
GSE7219_UNSTIM_VS_LPS_AND_ANTI_CD40_STIM_NIK_NFKB2_KO_DC_DN	3.4e-05	2.0e-03
GSE3982_CTRL_VS_LPS_48H_DC_DN	2.4e-04	7.5e-03

Figure S7: MSigDB over-representation analysis of DE genes identified in the LPS-dendritic cell dataset. Left panels (a)–(b) are based on pseudotime inferred by Slingshot; right panels (c)–(d) are based on pseudotime inferred by Monocle3-PI. **(a) & (c)** Numbers of MSigDB terms enriched (BH adjusted-$p < 0.01$) in the significant DE genes specifically found by PseudotimeDE or tradeSeq/Monocle3-DE in pairwise comparisons between PseudotimeDE and tradeSeq/Monocle3-DE in Fig. 4 (b) & (f). **(b) & (d)** Example MSigDB terms enriched in the Pseudotime-specific DE genes in (a) & (b). The explanation of terms can be found from MSigDB. All listed terms are related to the response of dendritic cells (DC) stimulated by LPS.

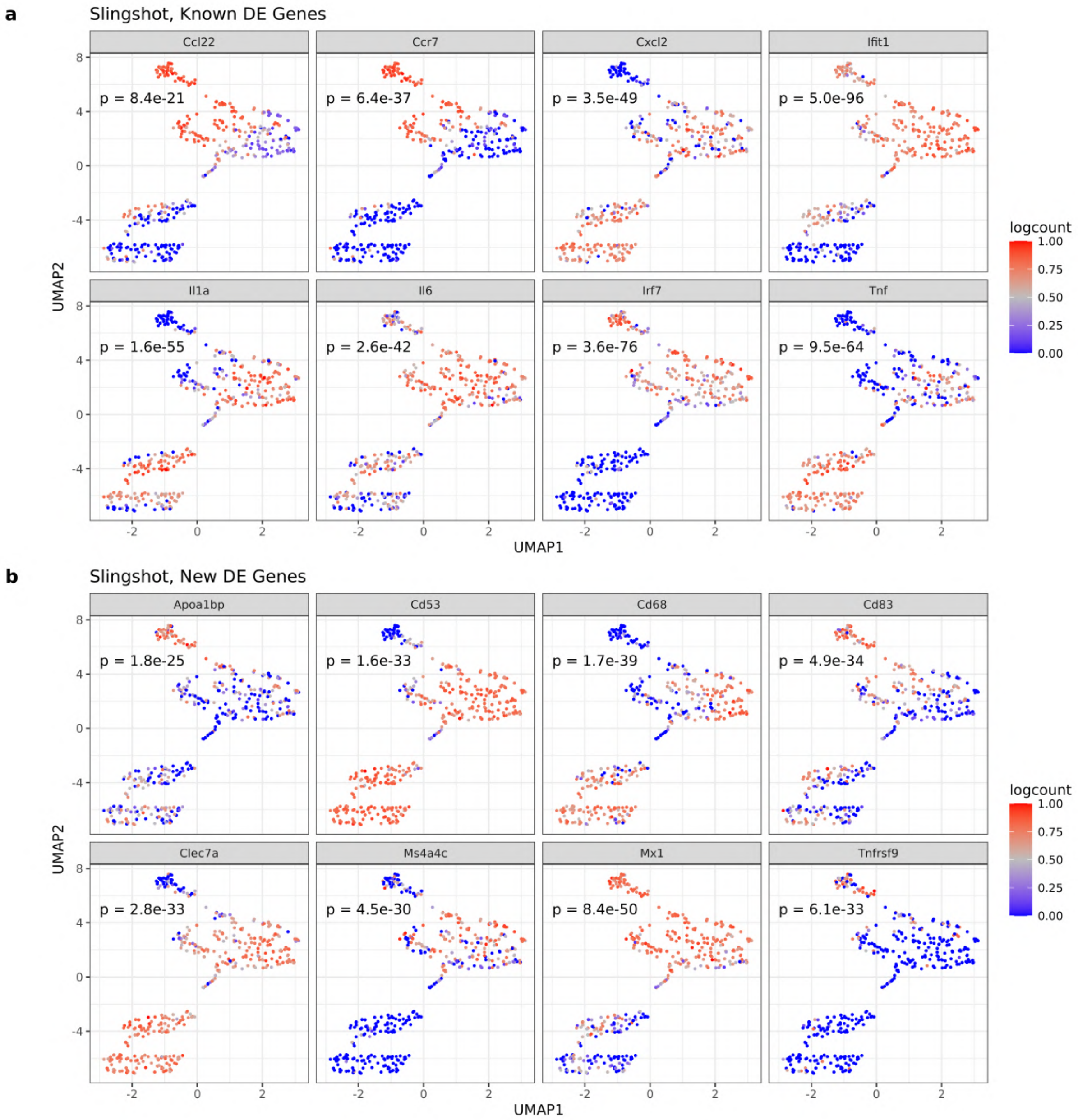


Figure S8: UMAP visualization of example DE genes identified by PseudotimeDE, using Slingshot as the pseudotime inference method, in the LPS-dendritic cell dataset. p-values returned by PseudotimeDE are reported for all the 16 example genes. **(a)** Examples of known DE genes, which have been reported as highly confident DE genes in the original study [31]. **(b)** Examples of new DE genes, which are identified by PseudotimeDE but found as non-DE by either tradeSeq or Monocle3-DE.

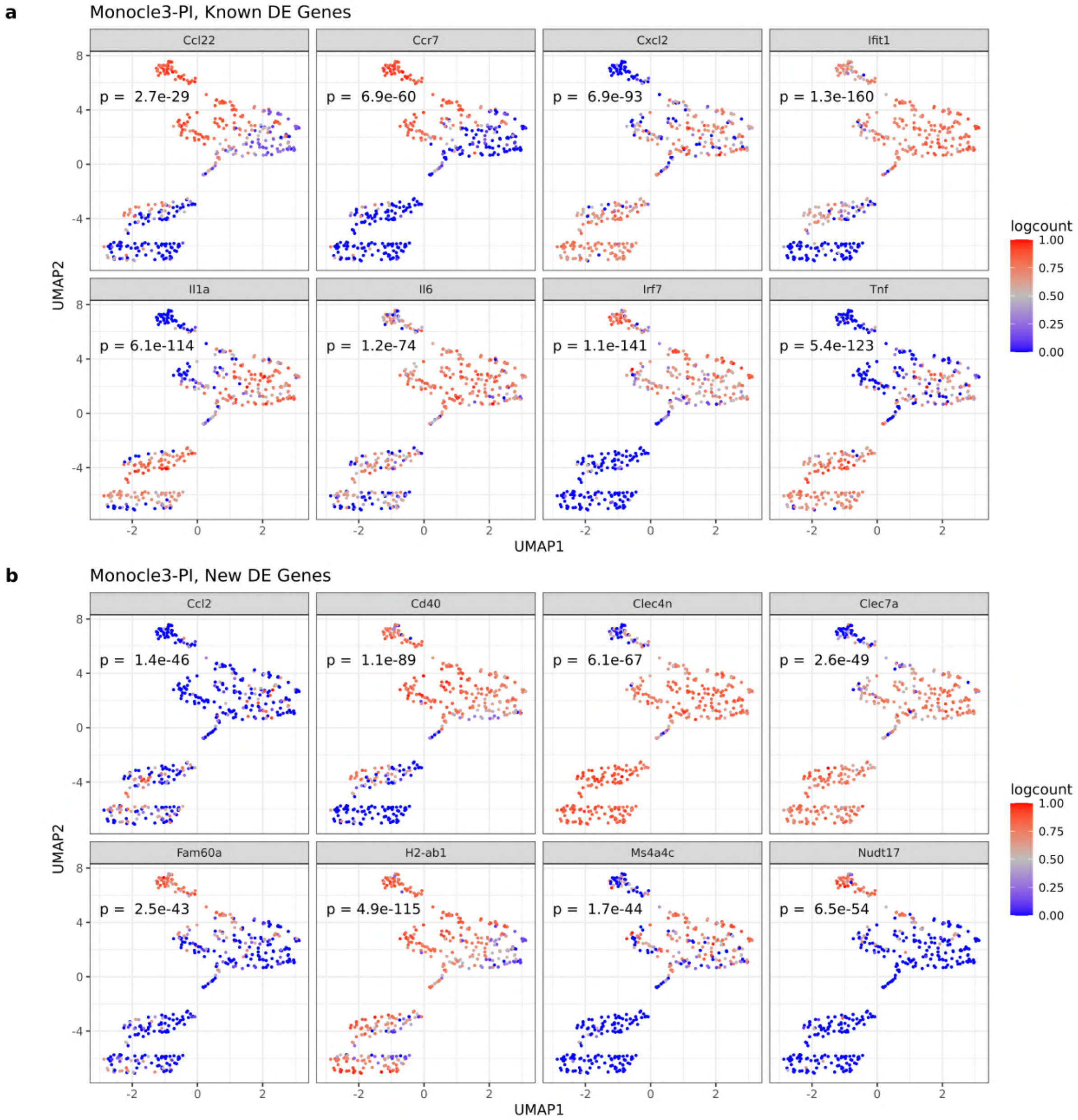


Figure S9: UMAP visualization of example DE genes identified by PseudotimeDE, using Monocle3-PI as the pseudotime inference method, in the LPS-dendritic cell dataset. p-values returned by PseudotimeDE are reported for all the 16 example genes. **(a)** Examples of known DE genes, which have been reported as highly confident DE genes in the original study [31]. **(b)** Examples of new DE genes, which are identified by PseudotimeDE but found as non-DE by either tradeSeq or Monocle3-DE.

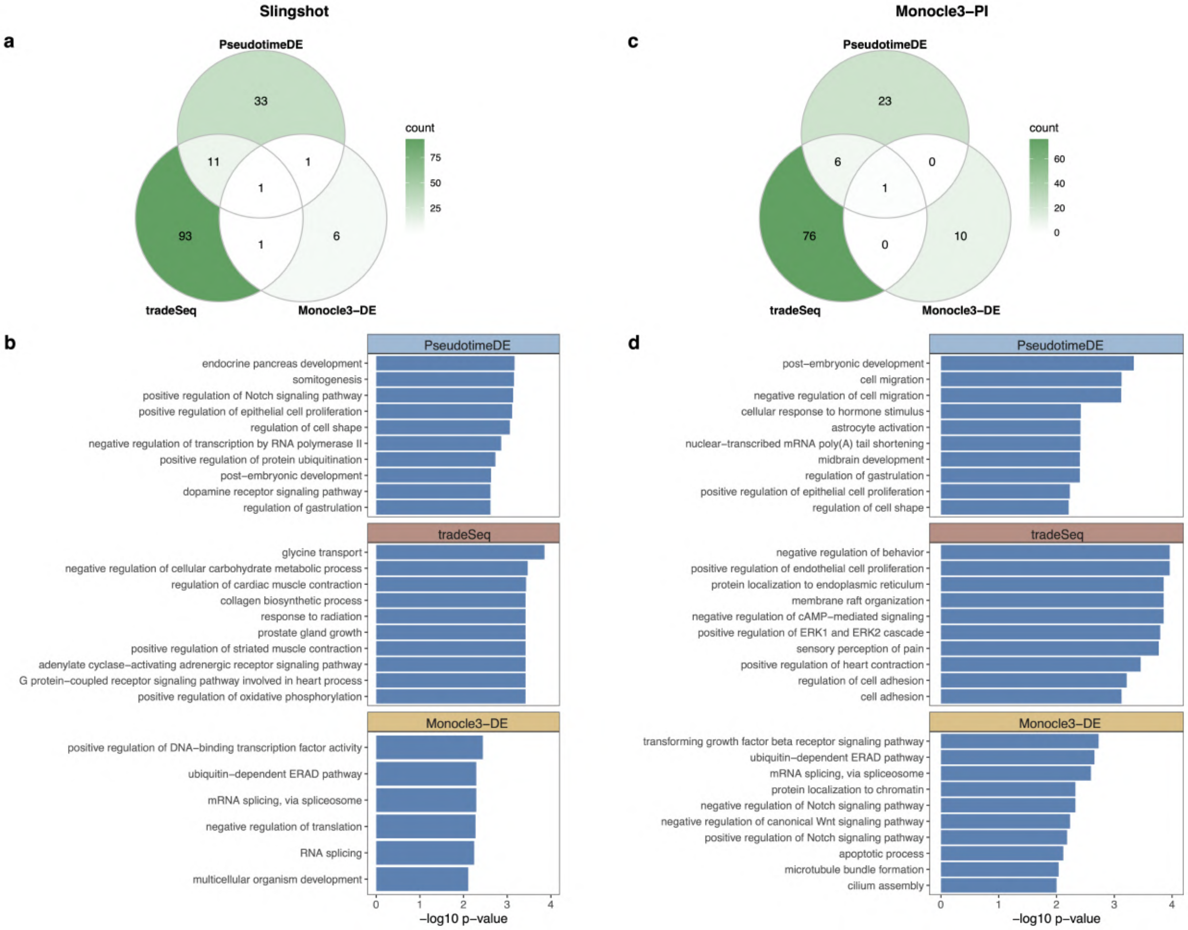


Figure S10: GO analysis of DE genes identified in the pancreatic beta cell maturation dataset. Left panels (a)–(b) are based on pseudotime inferred by Slingshot; right panels (c)–(d) are based on pseudotime inferred by Monocle3-PI. **(a) & (c)** Numbers of GO terms enriched ($p < 0.01$) in the significant DE genes found by each method in Fig. 5 (b) & (f). **(b) & (d)** Top 10 significant GO terms from each DE method.

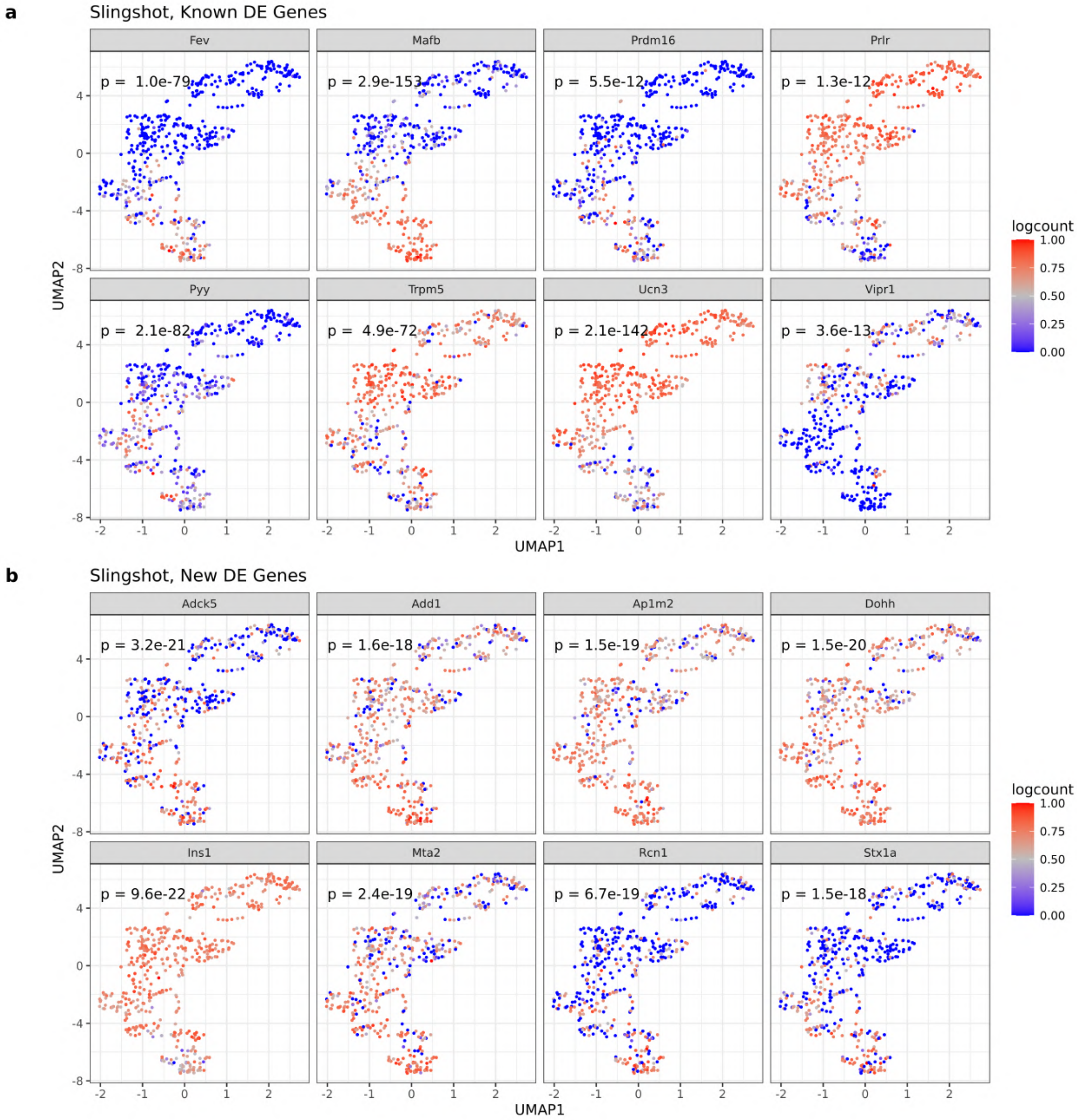


Figure S11: UMAP visualization of example DE genes identified by PseudotimeDE, using Slingshot as the pseudotime inference method, in the pancreatic beta cell maturation cell dataset. p-values returned by PseudotimeDE are reported for all the 16 example genes. **(a)** Examples of known DE genes, which have been reported as highly confident DE genes in the original study [33]. **(b)** Examples of new DE genes, which are identified by PseudotimeDE but found as non-DE by either tradeSeq or Monocle3-DE.

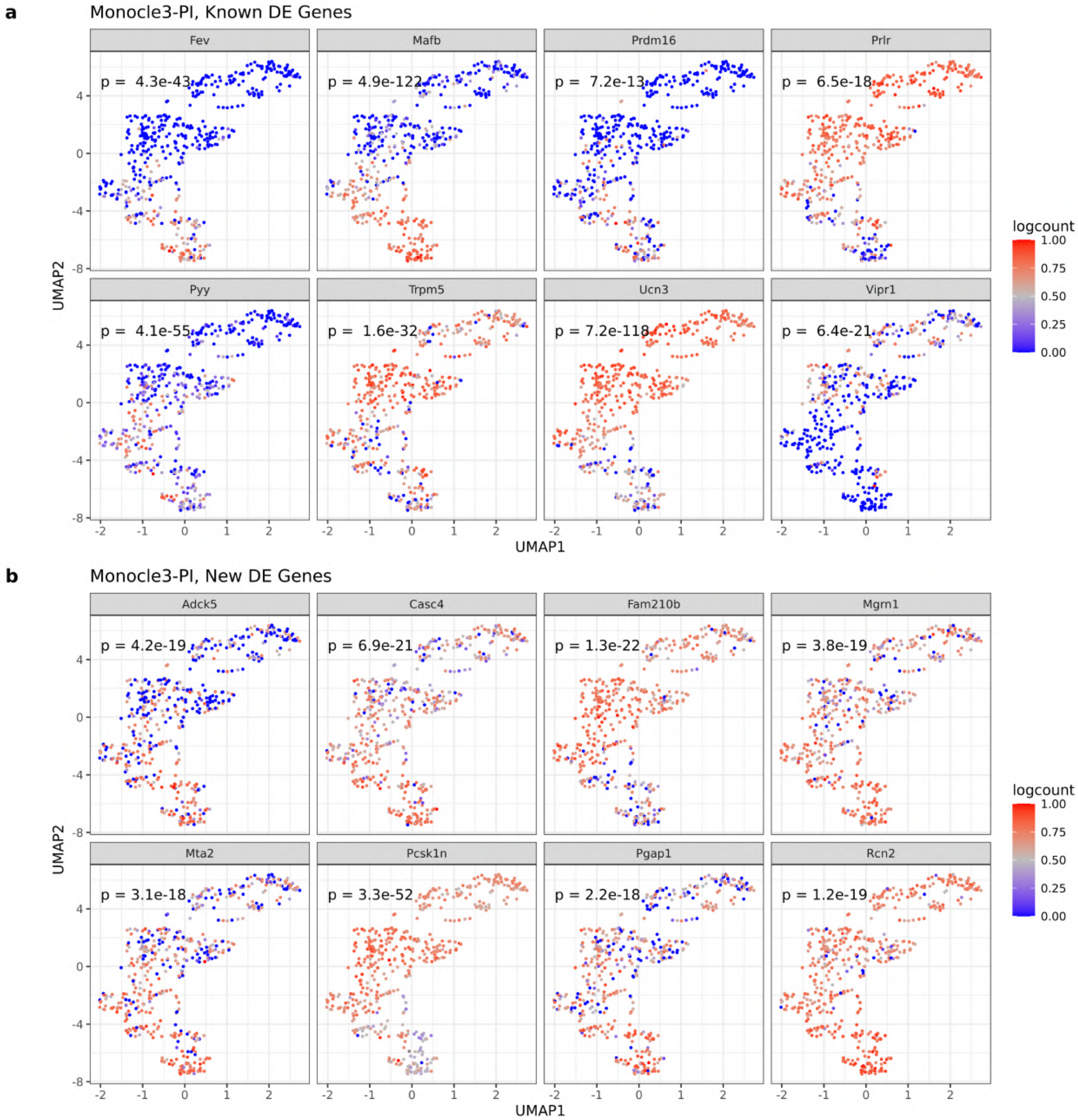


Figure S12: UMAP visualization of example DE genes identified by PseudotimeDE, using Monocle3-PI as the pseudotime inference method, in the pancreatic beta cell maturation cell dataset. p-values returned by PseudotimeDE are reported for all the 16 example genes. **(a)** Examples of known DE genes, which have been reported as highly confident DE genes in the original study [33]. **(b)** Examples of new DE genes, which are identified by PseudotimeDE but found as non-DE by either tradeSeq or Monocle3-DE.

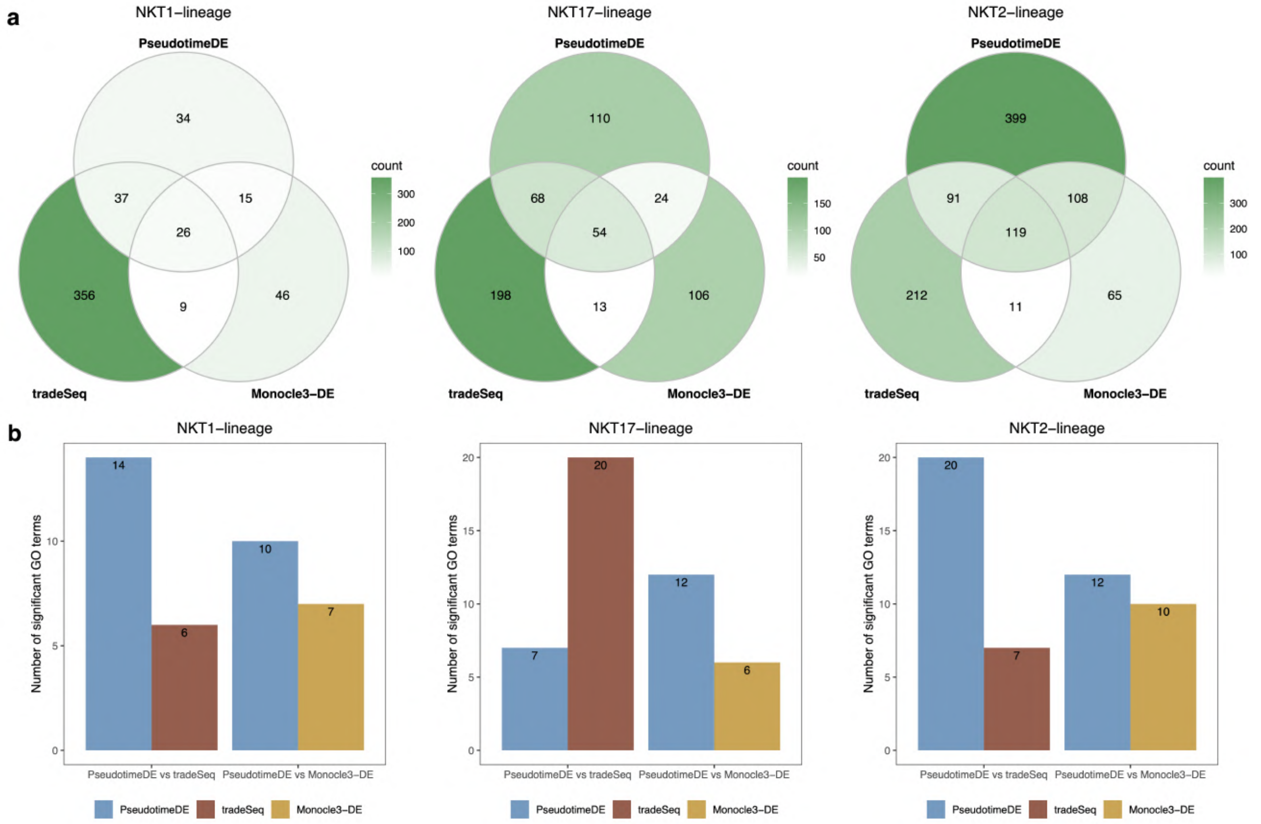


Figure S13: GO analysis of DE genes identified in the natural killer T cell dataset. **(a)** Venn plots showing the overlaps of the significant DE genes (BH adjusted-$p \leq 0.05$) identified by the three DE methods. **(b)** Numbers of GO terms enriched ($p < 0.05$) in the significant DE genes specifically found by PseudotimeDE or tradeSeq/Monocle3-DE in pairwise comparisons between PseudotimeDE and tradeSeq/Monocle3-DE in (a).

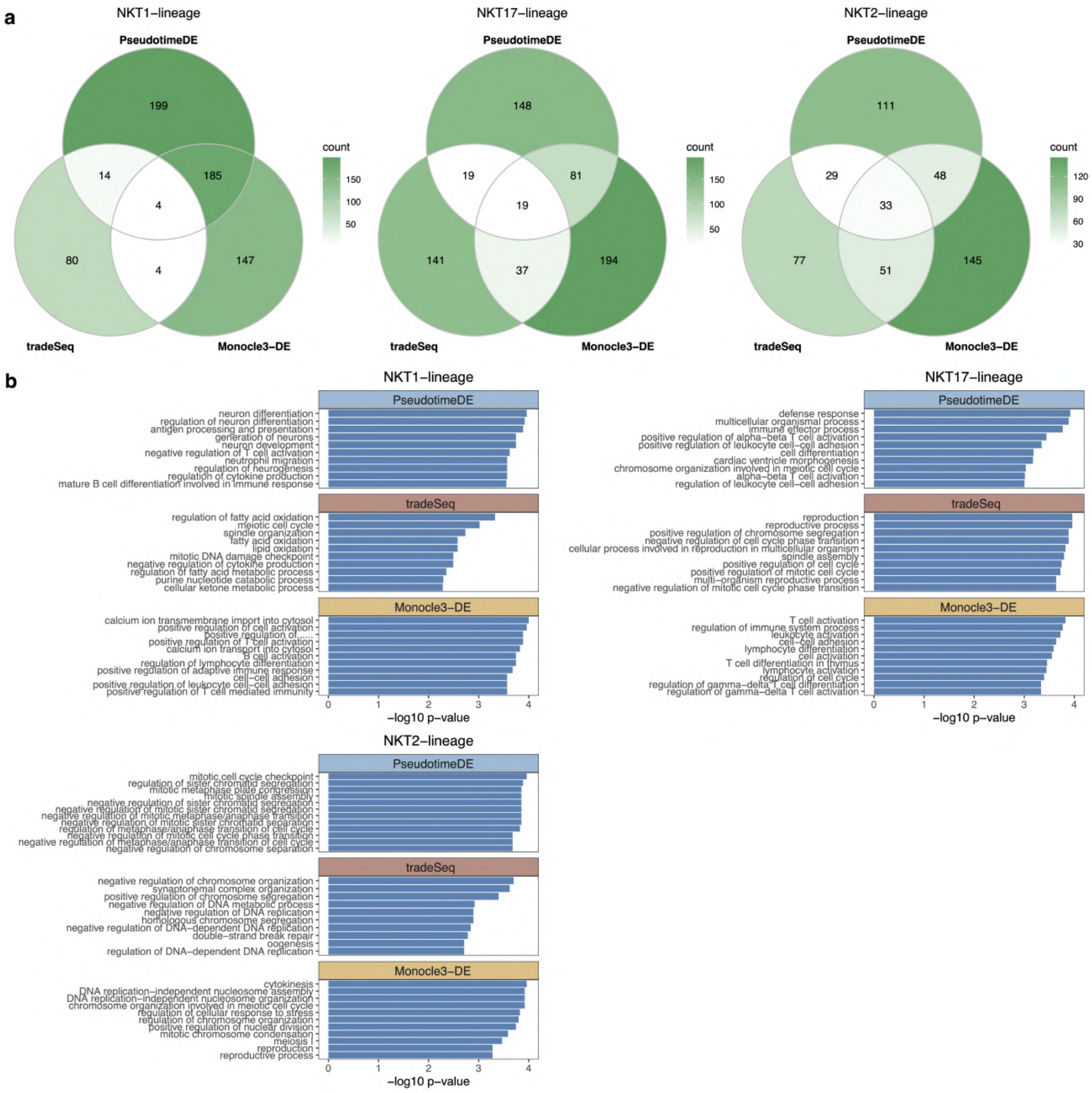


Figure S14: GO analysis of DE genes identified in the natural killer T cell dataset. (a) Numbers of GO terms enriched ($p < 0.05$) in the significant DE genes found by each method. **(b)** Top 10 enriched GO terms for each DE method.

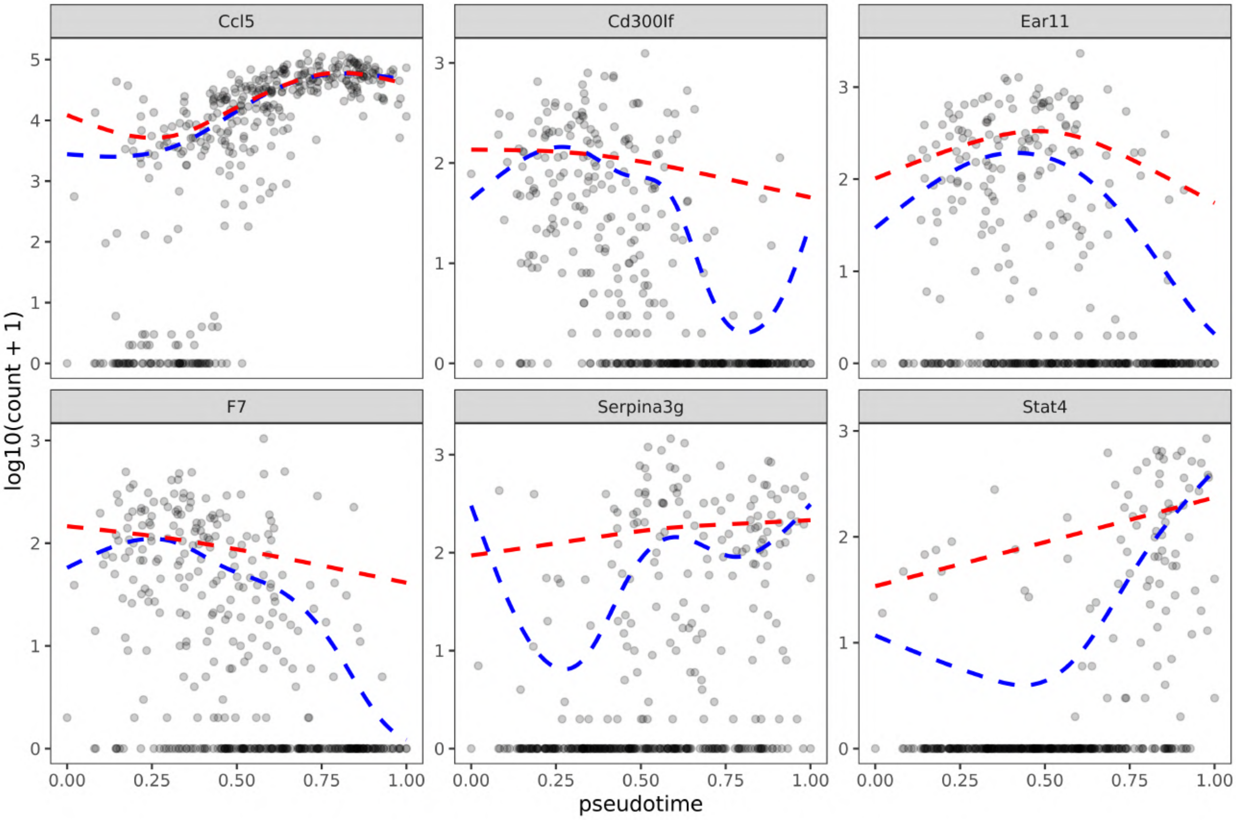


Figure S15: Comparison of NB-GAM and ZINB-GAM on the LPS-dendritic cell dataset with Slingshot pseudotime. Example fitted results of NB-GAM / ZINB-GAM on six genes from the LPS-dendritic cell dataset with pseudotime inferred by Slingshot. NB-GAM yields small p-values ($p < 1e - 10$) and ZINB-GAM yields large p-values ($p > 0.01$). Dashed blue lines and red lines are the fitted curves by NB-GAM and ZINB-GAM, respectively.

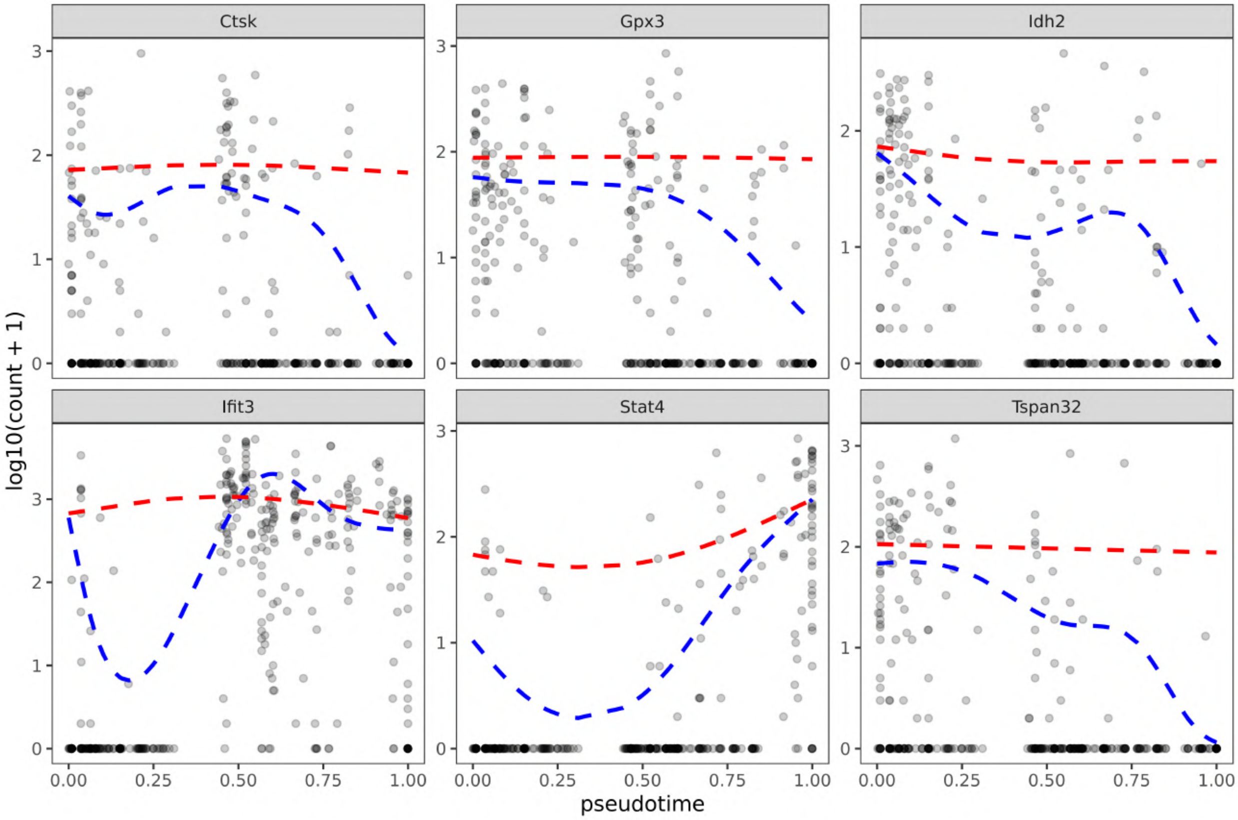


Figure S16: Comparison of NB-GAM and ZINB-GAM on the LPS-dendritic cell dataset with Monocle3-PI pseudotime. Example fitted results of NB-GAM / ZINB-GAM on six genes from the LPS-dendritic cell dataset with pseudotime inferred by Monocle3-PI. NB-GAM yields small p-values ($p < 1e - 10$) and ZINB-GAM yields large p-values ($p > 0.01$). Dashed blue lines and red lines are the fitted curves by NB-GAM and ZINB-GAM, respectively.

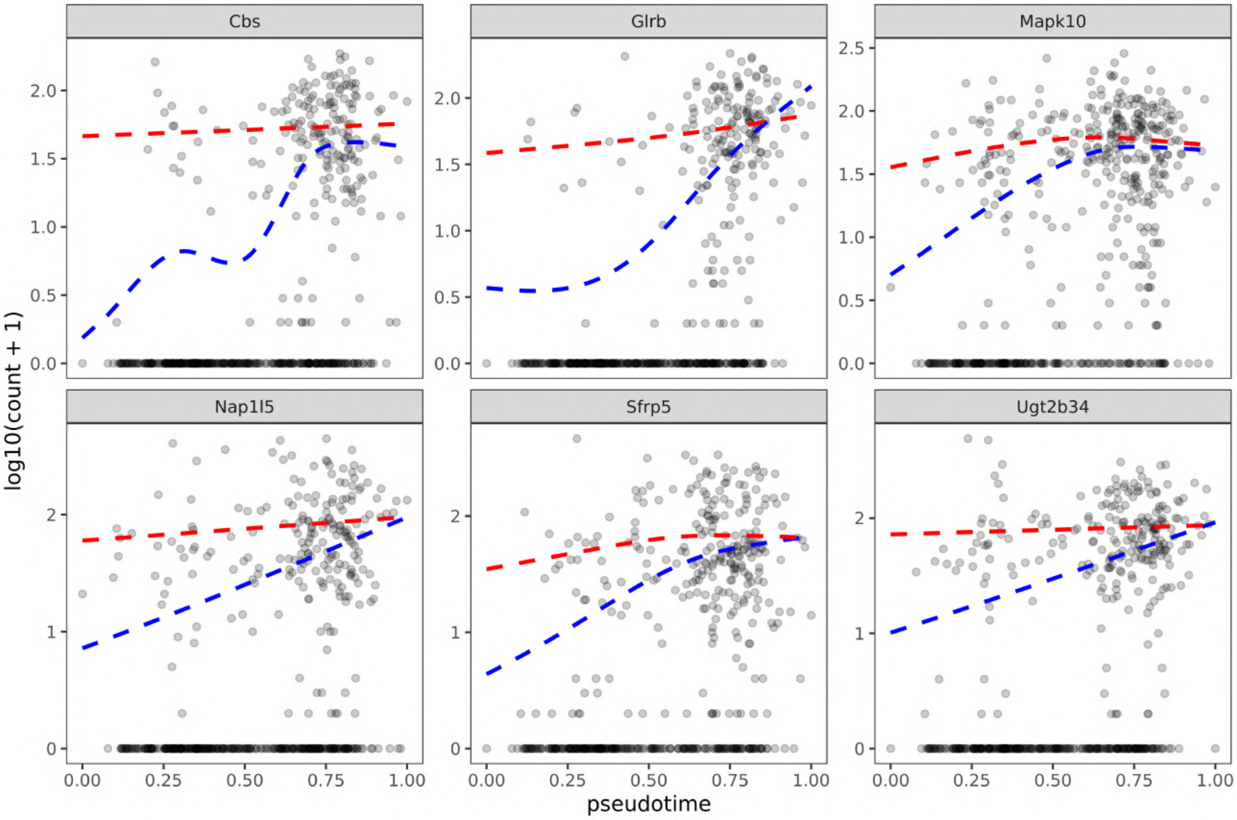


Figure S17: Comparison of NB-GAM and ZINB-GAM on the pancreatic beta cell maturation dataset with Slingshot pseudotime. Example fitted results of NB-GAM / ZINB-GAM on six genes fromthe pancreatic beta cell maturation dataset with pseudotime inferred by Slingshot. NB-GAM yields small p-values ($p < 1e-10$) and ZINB-GAM yields large p-values ($p > 0.01$). Dashed blue lines and red lines are the fitted curves by NB-GAM and ZINB-GAM, respectively.

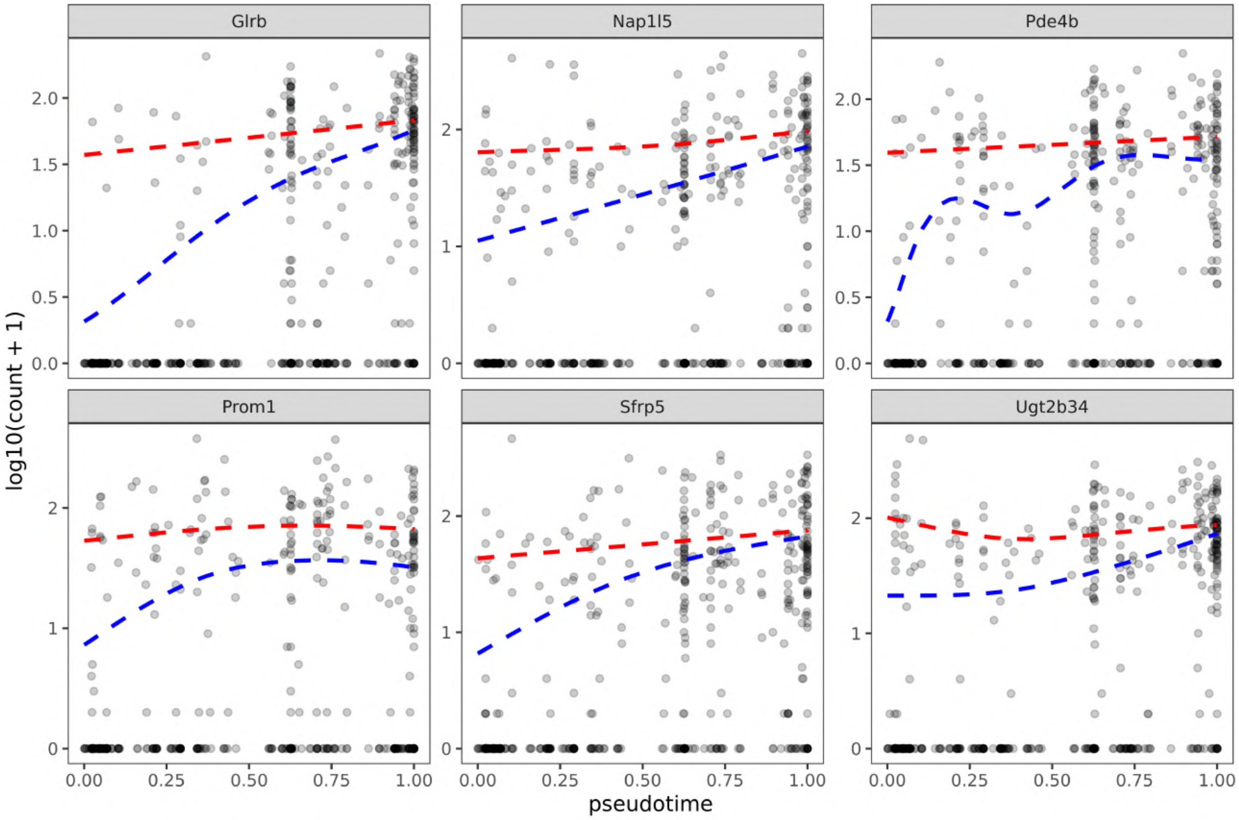


Figure S18: Comparison of NB-GAM and ZINB-GAM on the pancreatic beta cell maturation cell dataset with Slingshot pseudotime. Example fitted results of NB-GAM / ZINB-GAM on six genes from the pancreatic beta cell maturation with pseudotime inferred by Monocle3-PI. NB-GAM yields small p-values ($p < 1e-10$) and ZINB-GAM yields large p-values ($p > 0.01$). Dashed blue lines and red lines are the fitted curves by NB-GAM and ZINB-GAM, respectively.

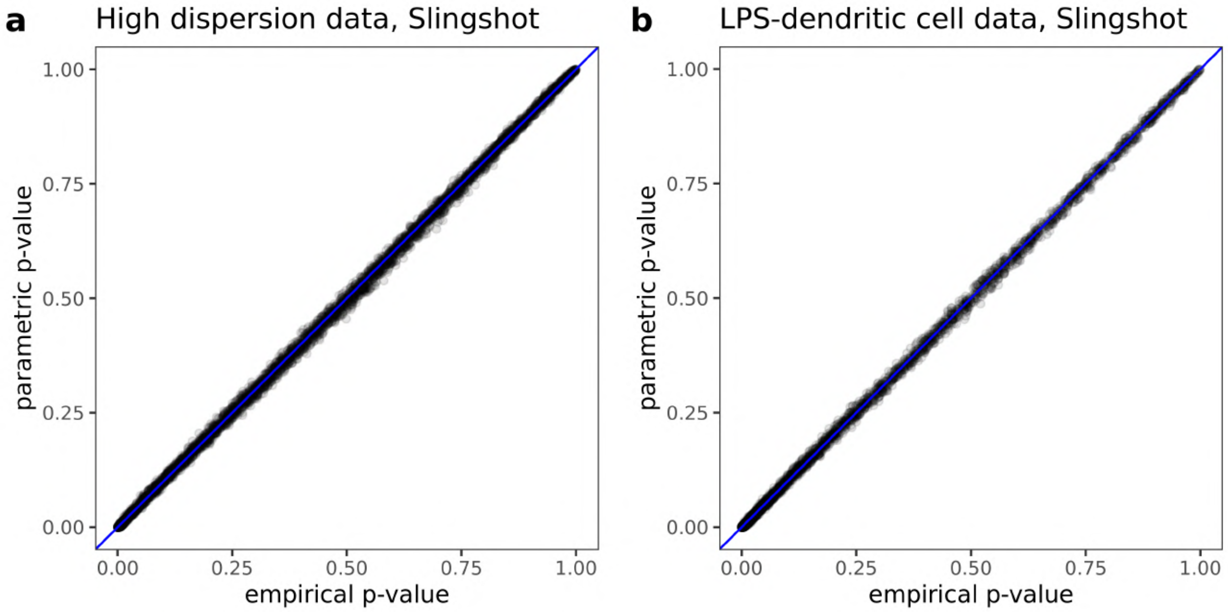


Figure S19: Comparison of empirical p-value and parametric p-value. Scatter plot of empirical p-values and parametric p-values (see **Methods**). **(a)** Scatter plot based on synthetic high dispersion dataset and pseudotime inferred by Slingshot. **(b)** Scatter plot based on LPS-dendritic cell dataset and pseudotime inferred by Slingshot. The parametric p-values are perfectly correlated with empirical p-values, suggesting that the parametric model well captures the estimated null distribution of test statistics.

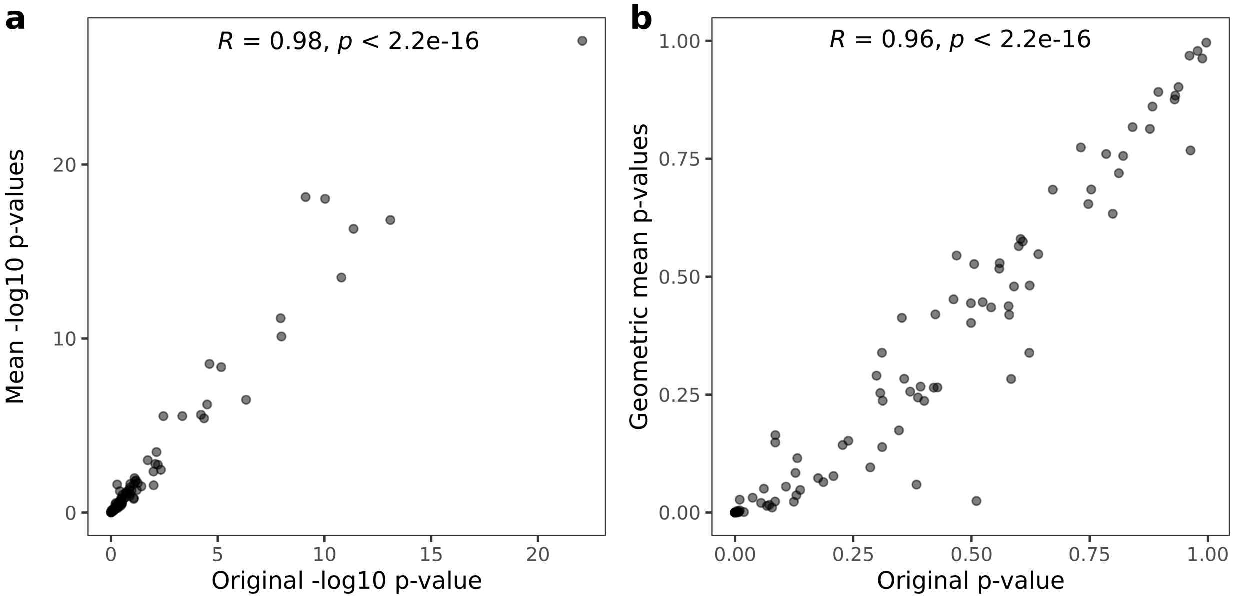


Figure S20: Comparison of p-values using 1000 subsamples and p-values using 100 subsamples. The p-values are based on based on synthetic high dispersion dataset and pseudotime inferred by Slingshot. **(a)** Scatter plot of the original p-value using 1000 subsamples on negative $-\log_{10}$ scale and the mean of 50 p-values using 100 subsamples on negative $-\log_{10}$ scale. The strict linearity (Pearson correlation coefficient $R = 0.98$) suggests that using 100 subsamples yield similar p-values as using 1000 subsamples on log scale. **(b)** Scatter plot of the original p-value using 1000 subsamples and the geometric mean of 50 p-values using 100 subsamples. The strict linearity (Pearson correlation coefficient $R = 0.96$) suggests that using 100 subsamples yield similar p-values as using 1000 subsamples on raw scale.

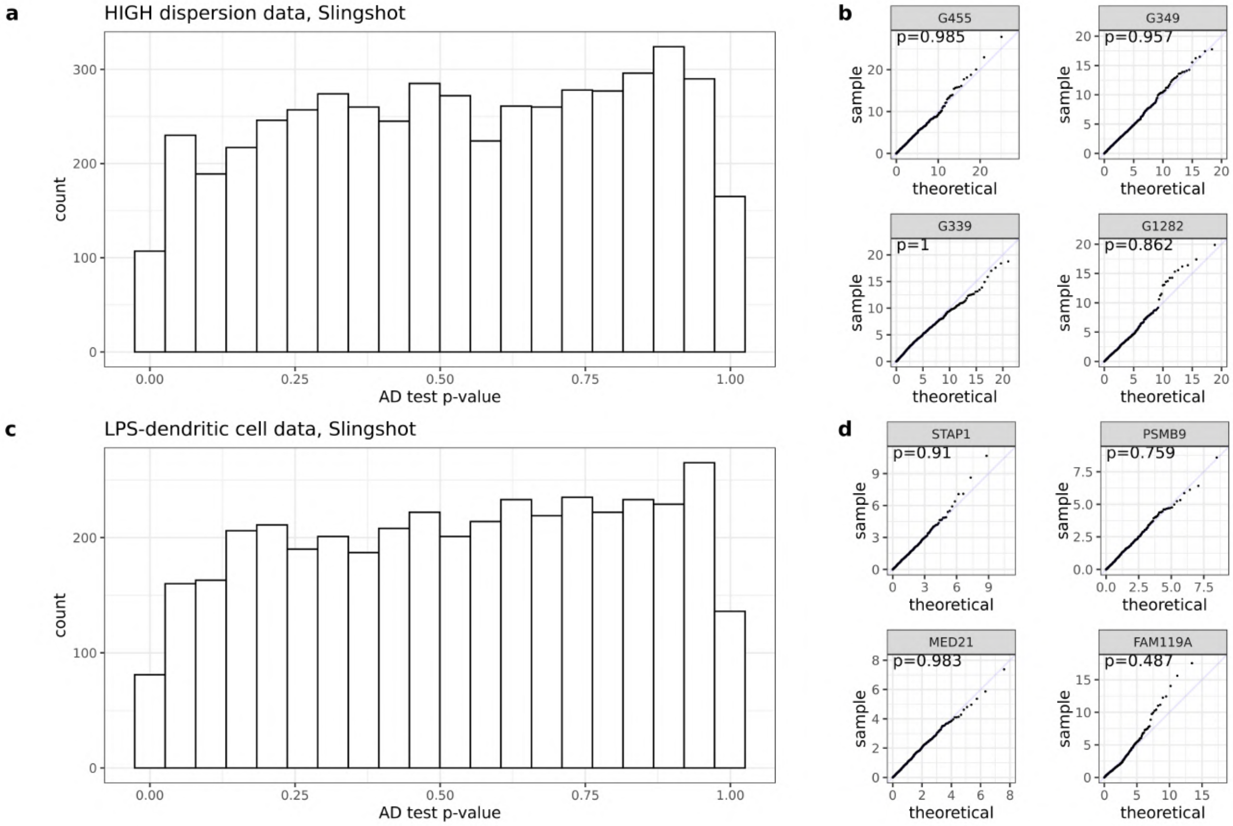


Figure S21: Goodness-of-fit of the parametric distribution. The p-values are from Anderson-Darling (AD) test, which measures the goodness-of-fit of the gamma/two-component gamma mixture distribution to the empirical null distribution generated by subsampling and permutation. **(a)** Histogram of AD test p-values based on the synthetic high dispersion dataset and pseudotime inferred by Slingshot. The distribution is approximately Uniform$[0, 1]$, indicating that the parametric distribution fits the empirical null distribution well. **(b)** Quantile-quantile plots comparing the empirical null distribution and its corresponding parametric fit for four random genes. **(c)** Histogram of AD test p-values based on the LPS-dendritic cell dataset and pseudotime inferred by Slingshot. The distribution is approximately Uniform$[0, 1]$, indicating that the parametric distribution fits the empirical null distribution well. **(d)** Quantile-quantile plots comparing the empirical null distribution and its corresponding parametric fit for four random genes.

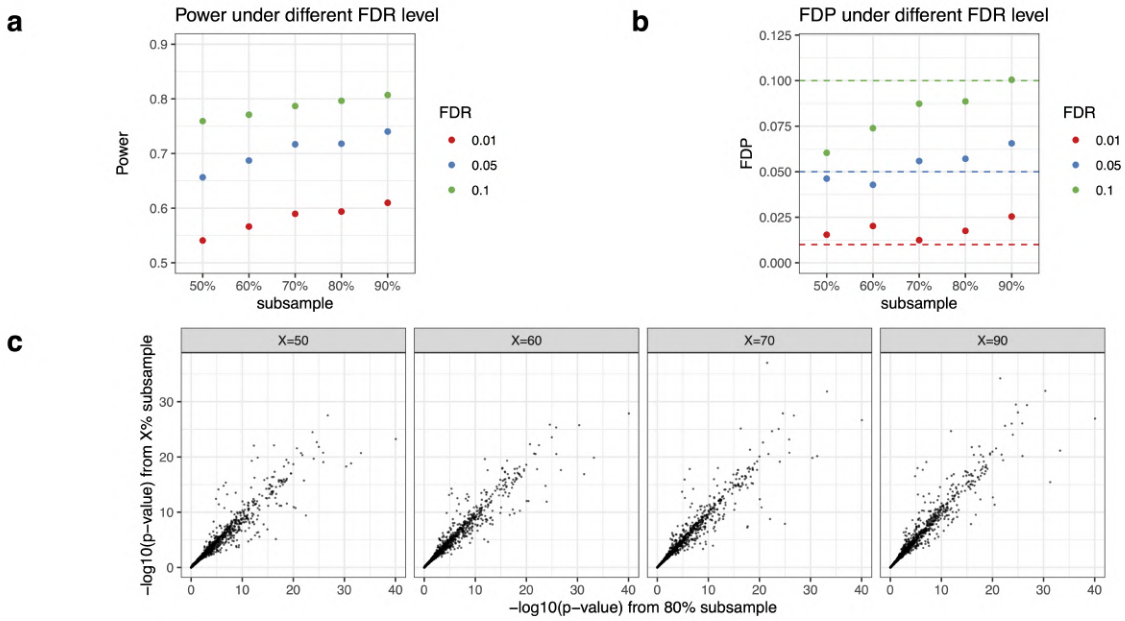


Figure S22: Robustness of PseudotimeDE to the subsampling proportion. Results are based on the synthetic high dispersion dataset and pseudotime inferred by Slingshot. **(a)** Power of PseudotimeDE using different subsampling proportions under FDR levels 0.01, 0.05, and 0.1. **(b)** FDP of PseudotimeDE using different subsampling proportions under FDR levels 0.01, 0.05, and 0.1. **(c)** Scatter plots of the default p-values using 80% as the subsampling proportion vs. the p-values using 50%, 60%, 70% or 90% as the subsampling proportion. The strong linearity (Pearson correlation coefficient $R \geq 0.96$) of p-values under different subsampling proportions confirms the robustness of PseudotimeDE to the subsampling proportion.