"output" — 2025/10/12 — 6:34 — page 1 — #1

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Gene expression

PseudotimeDE-fast: fast testing of differential gene expression along cell pseudotime

Yuheng Lai [1,†], Dongyuan Song [2,†], Lucy Xia [3,*] and Jingyi Jessica Li [4,5,*]

[1] Department of Statistics, University of Wisconsin Madison, [2] Department of Genetics and Genome Sciences, University of Connecticut Health Center, Farmington, CT 06030-6403, [3]Department of ISOM, School of Business and Management, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China, [4] Department of Statistics and Data Science, University of California, Los Angeles, CA 90095-1554, USA, [5] Biostatistics Program, Fred Hutchinson Cancer Center, Seattle, WA 98109, USA.

[†] These authors contributed equally to this work. *To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Identifying differentially expressed (DE) genes along cell pseudotime is crucial for understanding dynamic biological processes captured by single-cell RNA sequencing. However, existing DE methods either produce invalid p-values by ignoring the uncertainty in pseudotime inference or struggle to scale with the growing size of modern datasets. To address these limitations, we introduce PseudotimeDE-fast, a scalable method for detecting DE genes along pseudotime with well-calibrated p-values. Through comprehensive simulations and real-data analyses, we demonstrate that PseudotimeDE-fast delivers comparable or superior performance to existing approaches while offering substantial improvements in computational efficiency.

Availability: PseudotimeDE-fast is implemented in R with Rcpp acceleration and released under the MIT license. The source code is available at: `https://github.com/dsong-lab/PseudotimeDE`.

Contact: lijy03@fredhutch.org; lucyxia@ust.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) technologies have become a powerful tool for uncovering continuous transitions in cell populations. A common approach involves inferring a latent temporal variable, known as "pseudotime," from gene expression profiles to represent cells' relative positions along a developmental trajectory (Trapnell *et al.*, 2014). To interpret pseudotime, differential expression (DE) analysis is typically performed to identify genes with significant expression changes along the trajectory. Several methods have been developed for this purpose, such as tradeSeq (Van den Berge *et al.*, 2020), scMagSigPro (Srivastava *et al.*, 2024), and TDEseq (Fan *et al.*, 2024). However, these methods rely on regression models that treat pseudotime as fixed, ignoring the uncertainty in its inference. This oversight can lead to invalid p-values, as shown in prior studies (Campbell and Yau, 2016; Song and Li, 2021).

To consider the uncertainty in inferred pseudotime, we previously developed PseudotimeDE (Song and Li, 2021), the first DE method to explicitly account for this uncertainty. PseudotimeDE repeatedly performs trajectory (pseudotime) inference on subsampled cells and applies permutations to break the gene expression–pseudotime association, fitting a regression model to generate a null distribution of the test statistic. This approach yields well-calibrated p-values and good statistical power. However, its extensive computational demands, due to repeated model fitting on many subsamples, limit its scalability and broader adoption in the single-cell community.

To overcome the computational limitations of PseudotimeDE, we propose PseudotimeDE-fast, a novel method and updated R package for fast testing of gene expression changes along cell pseudotime. Unlike the methods that rely on regression models assuming fixed pseudotime, PseudotimeDE-fast tests the independence between gene expression and pseudotime by treating both as random variables. It implements a hypothesis test using a novel adaptation of the Bergsma–Dassios sign covariance τ^*—a robust extension of Kendall's tau—for sparse data, where $\tau^* = 0$ if and only if the two variables are independent (Bergsma and Dassios, 2014). Through comprehensive simulations and analysis of a large real dataset, we show that PseudotimeDE-fast produces well-calibrated p-values, achieves comparable or improved FDR control and power, and is over 100 times faster than existing methods.

2 Implementation

PseudotimeDE-fast is implemented in R and can be installed via `devtools::install_github("dsong-lab/PseudotimeDE")`. To address the computational bottleneck of its predecessor PseudotimeDE, it replaces the subsampling-and-permutation procedure with a direct, deterministic statistical test. Specifically, it reframes DE analysis as a formal test of independence between the pseudotime vector X and the expression vector Y_g of gene g.

The input consists of a scRNA-seq count matrix $\mathbf{Y} = [Y_1, \ldots, Y_p] \in \mathbb{R}^{n \times p}$, where n is the number of cells and p is the number of genes, and a pseudotime vector $X \in \mathbb{R}^n$ representing the inferred pseudotime of cells. For each gene $g \in \{1, \ldots, p\}$, PseudotimeDE-fast efficiently computes τ_n^*, a consistent estimator of the Bergsma–Dassios sign covariance τ^*:

$$\tau_n^*(X, Y_g) = \frac{(n-4)!}{n!} \sum_{\substack{1 \le i,j,k,l \le n \\ i,j,k,l \text{ distinct}}} a(X_i, X_j, X_k, X_l) \cdot a(Y_{gi}, Y_{gj}, Y_{gk}, Y_{gl}),$$

where

$$a(z_1, z_2, z_3, z_4) = \text{sign}\left(|z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4|\right).$$

The intuition for this measure, a powerful extension of the well-known Kendall's τ (Kendall, 1938), is that it moves beyond comparing simple pairs of points to evaluating all sets of four points (quartets). For each quartet, it checks whether the arrangement of points is "concordant" or "discordant" for both pseudotime and gene expression.

Previously, Heller and Heller (2016) introduced an algorithm to compute τ_n^* with $O(n^2)$ complexity, which becomes computationally prohibitive as n (the number of cells) increases. To address this, we developed an optimized algorithm that reduces the complexity of its core step to $O(Mn)$, where $M \ll n$ denotes the number of unique expression levels, often small due to sparsity in scRNA-seq data. Details are provided in **Supplementary Material** S1. Under the null hypothesis of independence between X and Y_g, τ_n^* admits a known limiting distribution, enabling efficient hypothesis testing (Nandy *et al.*, 2016). Compared to other rank-based independence tests with similar statistical properties (Shi *et al.*, 2022), our implementation achieves near-linear scalability for sparse data, while existing methods typically face computational bottlenecks.

3 Results

To evaluate the performance of PseudotimeDE-fast in terms of runtime, p-value validity, FDR control, and statistical power for detecting DE genes, we conducted simulations across varying numbers of cells (n) and used a large-scale real scRNA-seq dataset (Tsukui *et al.*, 2024). We compared PseudotimeDE-fast with state-of-the-art trajectory-based DE methods, including PseudotimeDE (Song and Li, 2021)—in both its asymptotic (fix) mode, which ignores pseudotime uncertainty and is not recommended, and its subsampling-and-permutation (permute) mode, which is accurate but computationally intensive - as well as tradeSeq (Van den Berge *et al.*, 2020) and TDEseq (Fan *et al.*, 2024). The details about the implementation and computational resources are described in **Supplementary Material** S2.

We generated synthetic datasets with $p = 2{,}000$ genes (20% DE) and varying numbers of cells $n \in \{1{,}000, 5{,}000, 10{,}000, 50{,}000, 100{,}000\}$ using scDesign3 (Song *et al.*, 2024), which was trained on a real scRNA-seq dataset of dentate gyrus neurogenesis (Hochgerner *et al.*, 2018). Fig. 1a shows results for four example genes: PseudotimeDE-fast reported highly significant p-values for three DE genes (*Ppia*, *Ncdn*, and *Calb2*) and an insignificant p-value for a non-DE gene (*Rab40b*).

Fig. 1b compares runtime across methods as n increases. All methods support multi-core parallelization, so we set the number of CPUs as 10 for every method. At $n = 10{,}000$, PseudotimeDE-fast completed in 124.29 seconds (CPU time): 298 times faster than tradeSeq, 348 times faster than PseudotimeDE-fix, 4,408 times faster than TDEseq, and over 24,013 times faster than PseudotimeDE-permute. With 10 cores, PseudotimeDE-fast finished in just 26.8 seconds. Note that TDEseq failed to finish within a reasonable runtime (48 hours) with $n = 50{,}000$ or more cells (**Supplementary Material** S2).

To assess p-value validity under the null, we compared p-values to the $\text{Uniform}[0, 1]$ distribution in two ways: (i) quantile-quantile (QQ) plots using $-\log_{10}$ p-values, and (ii) Kolmogorov–Smirnov tests using the raw p-values (Fig. 1c). PseudotimeDE-fast and PseudotimeDE-permute yielded well-calibrated p-values close to the expected uniform distribution. For DE gene detection at $n = 10{,}000$ (additional results in **Supplementary Fig.** S1), PseudotimeDE-fast achieved comparable power and FDR control to state-of-the-art methods while using far less computational time (Fig. 1d). For PseudotimeDE-fix, although its FDR was controlled, its p-values showed deviation from the expected uniform distribution (Fig. 1c; **Supplementary Fig.** S2). In addition, although PseudotimeDE-fast showed a slight power loss compared to PseudotimeDE-permute, the few missed genes were highly sparse and often of limited biological interest (**Supplementary Fig.** S3). These results highlight PseudotimeDE-fast as a scalable solution for large-scale pseudotime DE analysis. Note that this simulation has a high signal-to-noise ratio, so pseudotime can be estimated accurately and the "double-dipping" issue (Neufeld *et al.*, 2024) is relatively mild. If double-dipping remains a major concern, PseudotimeDE-fast may be combined with the synthetic-null-data approach employed by ClusterDE (Song *et al.*, 2025) to improve FDR control.

We further evaluated PseudotimeDE-fast using a large-scale scRNA-seq dataset of alveolar fibroblast lineage comprising $n = 35{,}096$ cells and $p = 12{,}834$ genes (Tsukui *et al.*, 2024). This dataset contains a single trajectory, and pseudotime was inferred using Slingshot (Street *et al.*, 2018). We applied PseudotimeDE-fast, PseudotimeDE-fix, and tradeSeq, which are the only feasible methods for this dataset, and excluded PseudotimeDE-permute and TDEseq due to scalability issues. PseudotimeDE-fast completed the analysis in under three hours, making it over 30 times faster than the other two methods, each of which required more than two days (Fig. 1e).

Since ground-truth DE genes are unknown, we assessed consistency across methods as a proxy for power. PseudotimeDE-fast identified a largely overlapping set of DE genes, sharing 63% with both other methods (Fig. 1f). Among DE genes missed by PseudotimeDE-fast but detected by both other methods (27%), 66.1% had zero expression in over 80% of cells, indicating high sparsity and limited informativeness. These results highlight that PseudotimeDE-fast offers substantial speed gains while maintaining comparable statistical power to existing approaches.

4 Discussion

Based on the Bergsma–Dassios sign covariance (an association measure for two random variables), PseudotimeDE-fast does not natively adjust for covariates such as batch effects or sequencing depth; users should therefore correct for confounders prior to analysis. Extending PseudotimeDE-fast to handle covariates would require a conditional (or partial) form of the Bergsma–Dassios sign covariance, which, to our knowledge, has not yet been developed and represents an interesting direction for future research.

5 Acknowledgements

The authors appreciate the comments and feedback from the members of the Junction of Statistics and Biology (https://jsb-lab.org). The authors thank Yuxin Yin for the help in refining the R package. The authors also thank Dr. Tatsuya Tsukui for sharing the dataset in Tsukui *et al.* (2024).

"output" — 2025/10/12 — 6:34 — page 3 — #3

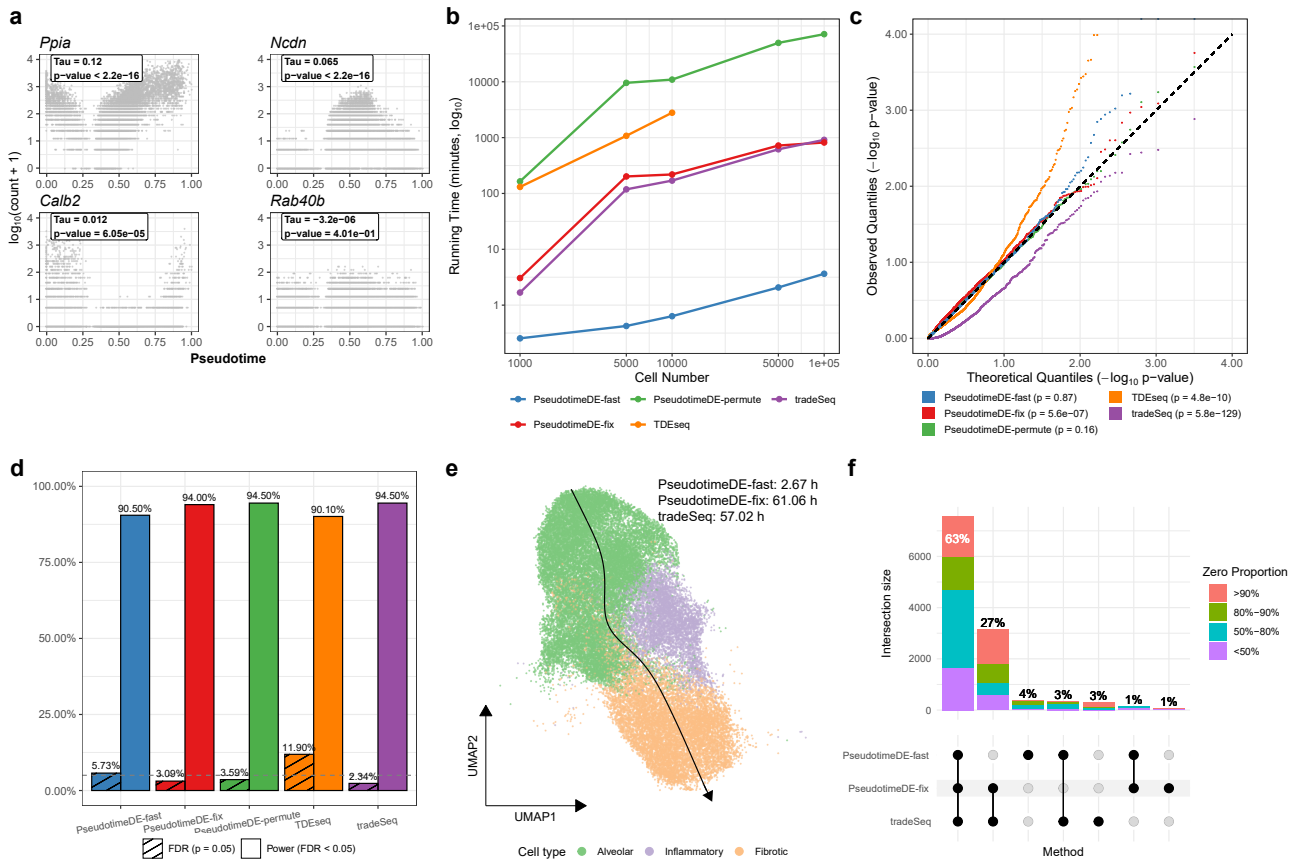

Fig. 1. Benchmarking PseudotimeDE-fast against other trajectory-based DE methods. (a) Expression of four example genes along pseudotime. The estimated τ_n^* and corresponding p-values are shown for each gene. (b) Runtime comparison across different cell numbers (n). PseudotimeDE-fast is significantly faster than all other methods. (c) Quantile-quantile plots of p-values under the null hypothesis on the $-\log_{10}$ scale. Only PseudotimeDE-fast and PseudotimeDE-permute produce well-calibrated p-values, with points falling along the diagonal and Kolmogorov-Smirnov test $p > 0.05$. (d) FDR and power in simulations. PseudotimeDE-fast achieves reasonable FDR control and comparable power to existing methods. (e) Application to the alveolar fibroblast lineage dataset (Tsukui et al., 2024). Cells are visualized by UMAP; colors denote cell types, and the curve indicates the inferred trajectory. PseudotimeDE-fast is 30× faster than PseudotimeDE-fix and tradeSeq. (f) UpSet plot showing overlaps in identified DE genes. PseudotimeDE-fast shares 63% of DE genes with both other methods, indicating high consistency.

6 Funding

This work was supported by National Science Foundation DBI-1846216 and DMS-2113754, NIH/NIGMS R01GM120507 and R35GM140888, Johnson and Johnson WiSTEM2D Award, Sloan Research Fellowship, and UCLA David Geffen School of Medicine W.M. Keck Foundation Junior Faculty Award (to J.J.L.).

References

Bergsma, W. and Dassios, A. (2014). A consistent test of independence based on a sign covariance related to kendall's tau. *Bernoulli*, pages 1006–1028.

Campbell, K. R. and Yau, C. (2016). Order under uncertainty: robust differential expression analysis using probabilistic models for pseudotime inference. *PLoS Computational Biology*, **12**(11), e1005212.

Fan, Y. *et al.* (2024). Powerful and accurate detection of temporal gene expression patterns from multi-sample multi-stage single-cell transcriptomics data with tdeseq. *Genome Biology*, **25**(1), 96.

Heller, Y. and Heller, R. (2016). Computing the bergsma dassios sign-covariance. *arXiv preprint arXiv:1605.08732*.

Hochgerner, H. *et al.* (2018). Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell rna sequencing. *Nature Neuroscience*, **21**(2), 290–299.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, **30**(1-2), 81–93.

Nandy, P. *et al.* (2016). Large-sample theory for the bergsma-dassios sign covariance. *Electronic Journal of Statistics*, **10**(2), 2287–2311.

Neufeld, A. *et al.* (2024). Inference after latent variable estimation for single-cell rna sequencing data. *Biostatistics*, **25**(1), 270–287.

Shi, H. *et al.* (2022). On the power of chatterjee's rank correlation. *Biometrika*, **109**(2), 317–333.

Song, D. and Li, J. J. (2021). Pseudotimede: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell rna sequencing data. *Genome Biology*, **22**(1), 124.

Song, D. *et al.* (2024). scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nature Biotechnology*, **42**(2), 247–252.

Song, D. *et al.* (2025). Synthetic control removes spurious discoveries from double dipping in single-cell and spatial transcriptomics data analyses. In *International Conference on Research in Computational Molecular Biology*, pages 400–404. Springer.

Srivastava, P. *et al.* (2024). scmasigpro: differential expression analysis along single-cell trajectories. *Bioinformatics*, **40**(7), btae443.

Street, K. *et al.* (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, **19**, 1–16.

"output" — 2025/10/12 — 6:34 — page 4 — #4

Trapnell, C. *et al.* (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, **32**(4), 381–386.

Tsukui, T. *et al.* (2024). Alveolar fibroblast lineage orchestrates lung inflammation and fibrosis. *Nature*, **631**(8021), 627–634.

Van den Berge, K. *et al.* (2020). Trajectory-based differential expression analysis for single-cell sequencing data. *Nature Communications*, **11**(1), 1201.