Statistical techniques for sparse compositional count data with applications to high-throughput single-cell RNA and amplicon sequencing

Johannes Ostner

München, 2024

Statistical techniques for sparse compositional count data with applications to high-throughput single-cell RNA and amplicon sequencing

Johannes Ostner

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

vorgelegt von
Johannes Ostner

München, den 09.12.2024

Erster Berichterstatter: Prof. Dr. Christian L. Müller (LMU München)
Zweiter Berichterstatter: Prof. Dr. Hongzhe Li (University of Pennsylvania)
Dritter Berichterstatter: Prof. Dr. Li Ma (Duke University)

Tag der Disputation: 08.05.2025

These days the statistician is often asked such questions as "Are you a Bayesian?" "Are you a frequentist?" "Are you a data analyst?" "Are you a designer of experiments?". I will argue that the appropriate answer to ALL of these questions can be (and preferably should be) "yes", and that we can see why this is so if we consider the scientific context for what statisticians do.

George E. P. Box

Acknowledgments

First of all, I want to thank Christian L. Müller for being an exceptional supervisor in every kind of way, reminding me time and time again that I chose the right path in pursuing a Ph.D. Thank you for your confidence in my abilities and also for patiently providing your invaluable expertise and guidance whenever I needed it. Also, I am grateful for the experiences I gained from presenting my work at numerous conferences and during my research stay, which were only possible due to your efforts. I could have never imagined a more supportive and encouraging supervisor.
I also want to thank Hongzhe Li for being a highly knowledgeable and encouraging member of my thesis committee, a welcoming and inspiring host during my research stay at UPenn, as well as an external reviewer for my thesis. Thank you to Li Ma for allowing me to give a presentation in his session at the ICSA symposium and agreeing to review my thesis as well. I would like to especially thank Benjamin Schubert, not only for giving me advice during my Ph.D. as a member of my thesis committee, but also for introducing me to the world of statistical genomics during my Master's thesis. Another special thank you goes to Maren Büttner for being a great collaborator and source of information, especially before and during the early stages of my Ph.D. Without the two of you, this chapter would have never even started. I would further like to thank all my other collaborators - especially Lukas Heumos for his efforts in integrating scCODA and tascCODA into the scverse, and Tim Kirk, Janne Gesine Thöming, and Adam Z. Rosenthal for supporting my work on bacterial scRNA-seq with data and biological insights.
Next, I would like to thank all current and former members of the bio-datascience lab. A special thank you to Mara Stadler for countless discussions about statistical and non-statistical topics, and for paving the way for me and all other Ph.D. students in this lab to come. Thank you to Roberto Olayo Alarcon, Oleg Vlasovets, Daniele Pugno, Viet Tran, Stefanie Peschel, and Medina Feldl for being amazing collaborators and colleagues, making this time unforgettable to me. Further thanks go to Fabian Schaipp, Luise Rauer, Aditya Mishra, Jinlong Ru, and Tong Wu. I would further like to thank Helmholtz Munich for financial support and the staff at the Computational Health Center, as well as the Statistics department at LMU.
I am also grateful to all people who have supported me during this time outside of my scientific life. Thank you to my colleagues and teammates at Verein Kulturleben in der Studentenstadt and VfL Waldkraiburg for keeping my social life in balance and allowing me to develop so many skills besides statistics. Thank you to all my friends for your unwavering support and encouragement during the past years and beyond. A very special thank you to Salomé Carcy and Anna Huber, who both played crucial roles during different periods of my Ph.D., as best friends to celebrate successes with as well as a shoulder to lean on during hard times. Finally, I would like to thank my grandfather Ernst Trautwein, my parents Michaela Trautwein-Ostner and Karl Ostner, my sisters Isabella Ostner and Lena Ostner, as well as all other members on both sides of my family. Words cannot describe how grateful I am for all of your help in overcoming any obstacle in my way and providing me with a safe haven at all times. This would not have been possible without all of you.

Zusammenfassung

Hochdurchsatz-Sequenzierungsverfahren (HDS) ermöglichen Forschenden, Typ und Funktion von großen Mengen an Zellen zu analysieren - entweder in ihrer Gesamtheit, zum Beispiel mithilfe von Amplicon-Sequenzierung in der Mikrobiomanalyse, oder individuell durch Einzelzellsequenzierung. Die primäre Datenstruktur für beide Technologien sind hochdimensionale Matrizen mit Zähldaten: Amplicon-Sequenzierung beschreibt typischerweise die Häufigkeit mikrobieller Taxa in jeder Stichprobe, während Einzelzellsequenzierung die Expressionslevel von Genen in jeder betrachteten Zelle liefert. Zusätzlich kann durch Bestimmung des Typs jeder Zelle in einem Einzelzellsequenzierungsexperiment eine Aggregation in eine Datenmatrix mit Stichproben und Zelltypen vorgenommen werden, welche denen der Amplicon-Sequenzierung ähnelt.

In beiden Fällen sind Veränderungen der Komposition unter demographischen, phänotypischen, oder umweltbezogenen Kovariaten von besonderem Interesse, wenngleich eine solche Analyse der diffenziellen Abundanz (DA) nicht trivial ist. HDS-Abundanzdaten enthalten oft mehr Komponenten als Stichproben, was besondere Vorsicht bei der Auswahl statistisch relevanter Effekte erfordert. Weiterhin stellen kleine Stichprobengrößen in der Einzelzellanalyse und dünnbesetzte mikrobielle Abundanzdaten weitere Herausforderungen für die Entwicklung geeigneter statistischer Methoden dar. Zuletzt rufen technische Einschränkungen eine Obergrenze in der Sequenzierungstiefe jeder Stichprobe hervor, was die Berücksichtigung kompositioneller Effekte notwendig macht.

Diese Dissertation besteht aus drei Abschnitten mit Artikeln, von denen jeder einen oder zwei Beiträge zur Kompositionsanalyse oder genereller statistischer Verarbeitung von HDS-Daten enthält. Demzuvor steht ein einleitender Teil, welcher die statistischen Grundlagen für die in den Artikeln verwendeten Methoden darlegt. Der erste Satz an Artikeln beschäftigt sich mit Bayesscher Modellierung und Tests zur differentiellen Abundanz in Hochdurchsatz-Sequenzierungsdaten und umfasst zwei Artikel. Im ersten Artikel wird `scCODA`, ein generatives Modell zur DA-Analyse von Zelltyp-Kompositionen aus der Einzelzellanalyse, vorgestellt. `scCODA` verwendet ein Dirichlet-Multinomialmodell zur Berücksichtigung der kompositionellen Bedingungen und führt Modellselektion mittels spike-and-slab Verteilungen und Schwellenwertsetzung auf deren Inklusionswahrscheinlichkeiten durch. Zusätzlich garantiert die automatische oder manuelle Auswahl einer Referenzkomponente volle Identifizierbarkeit. Die zweite Publikation beschäftigt sich mit den hierarchischen Strukturen von mikrobiellen Taxa und Zelltypen und erweitert `scCODA` um aggregierte Effekte auf den inneren Knoten des zugrundeliegenden Baumes. Dieses Modell, `tascCODA` genannt, werwendet spike-and-slab LASSO-Verteilungen und hierarchisch adaptive Regularisierungsstärken, um sich verändernde Komponenten und Gruppen von Komponenten zu identifizieren. Simulationsstudien und Anwendungen auf reelle Hochdurchsatz-Sequenzierungsdaten zeigen, dass `scCODA` und `tascCODA` die Falscherkennungsrate in Szenarien mit niedriger bis moderater Dimensionalität besser als vergleichbare Methoden kontrollieren und biologisch relevante Effekte erkennen.

Der zweite Abschnitt enthält ein Manuskript zu `cosmoDA`, eine Methode für DA-Tests auf HDS-Abundanzdaten unter der Berücksichtigung von Interaktionen zwischen den Komponenten. Durch die Modellierung von Kompositionsdaten durch a-b Power Interaction Modelle, eine Generalisierung der multivariaten Logit-Normalverteilung, kann `cosmoDA`

falsch positive Effekte, hervorgerufen durch paarweise Interaktion zwischen Komponenten, erkennen und vermeiden. Zusätzlich ermöglicht Score Matching-Optimierung effiziente Parameterschätzung des Modells, während regularisierte Schätzung der Interaktionen Identifizierbarkeit garantiert. Der Beitrag untersucht desweiteren die Möglichkeit zur Vermeidung der Imputation von Nulleinträgen durch die Verwendung von Box-Cox-Transformationen im Zusammenhang mit der a-b Power Interaction-Modellfamilie.
Die Artikel im letzten Teil der Arbeit definieren beste Verfahrensweisen für die Analyse von Einzelzell-Sequenzierungsdaten. Der erste Beitrag stellt ein Verfahren für die automatisierte statistische Verarbeitung von bakteriellen Einzelzellsequenzierungsdaten mit Namen `BacSC` vor. Das Verfahren kombiniert Ideen aus dem data thinning und Vergleiche mit negativen Kontrolldaten, um die Selektion von Hyperparametern zur Dimensionsreduktion, Visualisierung und Gruppierung zu automatisieren, sowie die Falscherkennungsrate unter "double dipping"-Bedingungen in der differentiellen Genexpressionsanalyse zu kontrollieren. `BacSC` berücksichtigt weiterhin die extreme Nullinflation und geringe Sequenzierungstiefe bakterieller Einzelzellsequenzierungsdaten während der Varianzstabilisierung und zeigt Verbesserungen bei der Generierung von Nulldaten unter diesen Bedingungen. Der zweite Beitrag in diesem Abschnitt definiert beste Vorgehensweisen und Beispielanalysen für Forscher bei der Kompositionsanalyse von Einzelzellsequenzierungsdaten mit `scCODA` und `tascCODA`.

Summary

High-throughput sequencing (HTS) methods enable researchers to analyze the type and function of large numbers of cells either in bulk, for example by amplicon sequencing for microbiome analysis, or individually through single-cell RNA sequencing (scRNA-seq). The primary data structures for both technologies are high-dimensional count matrices: Amplicon sequencing data typically describes the abundance of microbial taxa in each sample, while scRNA-seq yields expression counts of genes for each of the sequenced cells. Additionally, determining the type of each cell in a scRNA-seq experiment with multiple samples allows aggregation into a sample by cell-type count matrix, similar to amplicon sequencing.

In both cases, changes in the feature composition under demographic, phenotypical, or environmental covariates are of particular interest, but such differential abundance (DA) analysis is not straightforward from a statistical perspective. HTS abundance datasets often contain more features than samples, warranting specific care in the selection of statistically relevant effects, while low sample sizes in scRNA-seq and high sparsity in microbial abundance data pose further challenges for the development of suitable statistical methods. Finally, technological limitations induce an upper bound on the sequencing depth for each sample, which makes accounting for compositional effects a necessity.

This dissertation comprises three areas of articles, each providing one or two contributions in compositional analysis or general statistical processing of HTS data. They are preceded by an introductory part detailing the statistical foundations for the methods used throughout the contributions. The first section of articles is concerned with Bayesian modeling and differential abundance testing of high-throughput sequencing data and consists of two articles. In the first contribution, `scCODA`, a generative model for DA testing of cell-type compositions from scRNA-seq, is introduced. `scCODA` uses a Dirichlet-Multinomial model to account for the compositional constraints and performs model selection through spike-and-slab priors and thresholding on the posterior inclusion probability. Additionally, the automatic or manual selection of a reference feature ensures full identifiability of the model. The second publication notes the hierarchical structure of microbial taxa and cell-types alike and extends `scCODA` to consider aggregated effects on the nodes of the underlying feature tree. The resulting model, called `tascCODA`, utilizes spike-and-slab LASSO priors and hierarchically adaptive regularization penalties to find differentially abundant features and groups of features over the entire tree. Simulation studies and applications to scRNA-seq data show that `scCODA` and `tascCODA` have better FDR control than other DA testing methods in low- to moderate sample-size settings and select biologically relevant effects.

The second section contains a manuscript on `cosmoDA`, a method for DA testing of HTS abundance data in the presence of feature-feature correlations. By modeling compositional data through a-b power interaction models, a generalization of the multivariate logistic normal distribution, `cosmoDA` detects and avoids spurious effects caused by first-order associations between features. In addition, score matching optimization allows for very efficient parameter estimation of the proposed model, while penalized estimation of the interaction matrix ensures model identifiability. The contribution further examines the use of Box-Cox transformations in conjunction with the a-b power interaction model

family to eliminate the need for zero imputation in compositional data.
The manuscripts in the final part of the thesis define best practices for the analysis of scRNA-seq data. The first contribution provides a framework for automatic statistical processing of gene expression data from single-cell sequencing on bacteria, called `BacSC`. The pipeline combines ideas from data thinning and comparisons with negative control data to automate the selection of hyperparameters for dimension reduction, visualization, and clustering, and guarantees FDR control under "double dipping" conditions in differential gene expression testing. `BacSC` further accounts for the extreme zero inflation and low sequencing depth of bacterial scRNA-seq data during variance stabilization and presents improvements to null data generation under these conditions. The second contribution in this section provides best practices and example workflows for researchers when performing compositional analysis of scRNA-seq data with `scCODA` and `tascCODA`.

List of contributed publications

This thesis is based on the following publications and manuscripts (listed in chronological order):

1. Büttner, M.*, **Ostner, J.***, Müller, C. L., Theis, F. J., and Schubert, B. (2021). scCODA is a Bayesian model for compositional single-cell data analysis. *Nat. Commun. 12, 6876.* doi: https://doi.org/10.1038/s41467-021-27150-6

   * joint first co-authorship

   (See also [1] in the bibliography)

2. **Ostner, J.**, Carcy, S., and Müller, C. L. (2021). tascCODA: Bayesian Tree-Aggregated Analysis of Compositional Amplicon and Single-Cell Data. *Front. Genet. 12, 766405.* doi: https://doi.org/10.3389/fgene.2021.766405

   (See also [2] in the bibliography)

3. **Ostner, J.**, Li, H., and Müller, C.L. (2024). Score matching for differential abundance testing of compositional high-throughput sequencing data. *bioRxiv, 2024-12.* doi: https://doi.org/10.1101/2024.12.05.627006

   (See also [3] in the bibliography)

4. **Ostner, J.**, Kirk, T., Olayo-Alarcon, R., Thöming, J., Rosenthal, A. Z., Häussler, S., et al. (2024). BacSC: A general workflow for bacterial single-cell RNA sequencing data analysis. *bioRxiv, 2024-06.* doi: https://doi.org/10.1101/2024.06.22.600071

   (See also [4] in the bibliography)

5. Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., [...] **Single-cell best practices consortium***, Schiller, H.B., Theis, F.J. (2023). Best practices for single-cell analysis across modalities. *Nat. Rev. Genet. 24, 550–572.* doi: https://doi.org/10.1038/s41576-023-00586-w

   * I am part of the Single-cell best practices consortium

   (See also [5] in the bibliography)

Other publications not included in this thesis:

6. Carcy, S., **Ostner, J.**, Tran, V., Menden, M., and Müller, C. L. (2024). MetaIBS - large-scale amplicon-based meta analysis of irritable bowel syndrome. *bioRxiv, 2024-01* doi: https://doi.org/10.1101/2024.01.22.575775

7. Heumos, L., Ji, Y., May, L., Green, T., Zhang, X., Wu, X., [...], **Ostner, J.**, [...], Theis, F.J. (2024). Pertpy: an end-to-end framework for perturbation analysis. *bioRxiv, 2024-08.* doi: https://doi.org/10.1101/2024.08.04.606516

Contents

1. Introduction

High-throughout sequencing (HTS) experiments enable the dissection of a heterogeneous population of cells into its individual types, giving new insights into the cellular constituents of a sample [6, 7]. Two of the most popular HTS methods are amplicon sequencing [8] and single-cell RNA sequencing (scRNA-seq) [9]. While the former method is widespread in the analysis of microbial communities, the latter is primarily used on eukaryotic tissues, but has recently been successfully adapted to bacteria as well [10]. Besides careful collection, treatment, and processing of the physical samples, correct statistical analysis of the resulting sequencing data is essential to obtain meaningful biological insights [11]. HTS results are often aggregated into count matrices $\tilde{\boldsymbol{X}} \in \mathbb{N}_0^{n \times p}$, detailing the abundance of p features (e.g. eukaryotic cell types or microbial taxa) for n samples.
A fact that is often overlooked in HTS datasets is their nature as a collection of size-constrained snapshots from larger populations (Figure 1.1). Due to limitations on the amount of cells that can be analyzed in a single sequencing run, the total number of sequenced cells in a sample is not related to the number of cells in the original tissue or environment [12, 13, 14]. Despite their apparent count structure, HTS abundance data should therefore be seen as compositional data, detailing only proportions instead of absolute values. Following Aitchison's foundational definition [15], p-dimensional compositional data vectors possess a natural sum-to-one constraint and are thus located on the $(p-1)$-dimensional probability simplex

$$\Delta_{p-1} = \left\{ \boldsymbol{x} \in \mathbb{R}^p : \boldsymbol{x} \succeq \mathbf{0},\, \mathbf{1}_p^\top \boldsymbol{x} = 1 \right\}. \tag{1.1}$$

To transform the absolute abundances $\tilde{\boldsymbol{X}}$ into relative abundance data $\boldsymbol{X}$ located on the simplex, the counts in each sample must be divided by their sum over all features. This step is known as the closure operation (Figure 1.1):

$$\boldsymbol{X}_i = C(\tilde{\boldsymbol{X}}_{\boldsymbol{i}}) = \frac{\tilde{\boldsymbol{X}}_{\boldsymbol{i}}}{\sum_{j=1}^{p} \tilde{X}_{i,j}} \; \forall i = 1 \dots n. \tag{1.2}$$

The sum-to-one constraint makes the simplex a bounded and thus non-euclidean space, invalidating many classical statistical principles [15]. Most importantly, the features in compositional datasets possess an inherent negative correlation [16], which necessitates joint analysis of all features instead of assuming feature independence.
One central task in the analysis of HTS abundance data is its description through generative statistical models, which can not only serve as a tool to accurately depict the data distribution, but also help to elucidate the underlying biological mechanisms and principles [11]. In larger studies with samples collected from many subjects or environments, the impact of sample-specific metadata ($\boldsymbol{Y}$ in Figure 1.1) on the feature composition is of particular interest. This metadata can include clinical covariates such as disease status

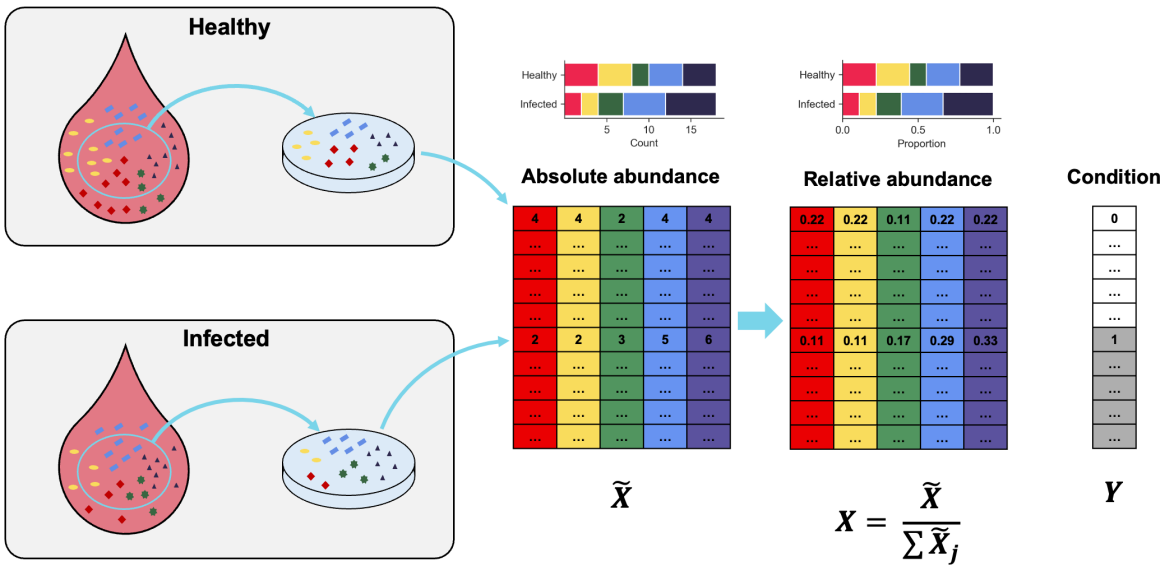


Figure 1.1.: **Schematic representation of the HTS data generation process.** Representative samples (Here: Blood cells) are collected from different environments. Then, the number of occurrences of each feature (Here: Cell types) in each sample is determined through HTS. To allow for sample comparison, the relative abundance must be considered.

and treatment group, information about the subject (age, sex, diet, . . .), or environmental factors like temperature, pH value, or collection location. Here, generative models can serve as a flexible and interpretable foundation to describe a plethora of experimental scenarios and designs. In addition to modeling the impact of covariates on the feature composition, the detection of statistically relevant effects is essential, as it reveals important associations between external factors and the individual components of the cell population. If the covariate is binary, i.e. when evaluating the compositional differences between two groups, such analysis is known as differential abundance (DA) testing [8, 17]. The most common example is the determination of compositional differences between a control group and group with a disease or a specific treatment. The natural anticorrelation between the features plays an essential role when modeling changes in compositional data, as a shift in the relative abundance of one feature induces changes in the relative abundance of all other features due to the sum-to-one constraint of the simplex. Ignoring this property during DA testing can quickly lead to false positive associations [1, 12, 18].

1.1. Compositional Data in Genomics

To gain a better understanding for the characteristics of different kinds of high-throughput sequencing abundance data and resulting challenges for generative modeling, it is

helpful to take a closer look at the data generation process. Over the last 20 years, the field of molecular biology was revolutionized by rapid improvements in sequencing technologies, allowing the parallel analysis of millions of RNA sequences, also known as transcriptomes, at ever-decreasing costs [19, 20]. These next-generation sequencing (NGS) methods generally follow an experimental pattern consisting of library preparation, sequencing, and analysis (Figure 1.2A, B). Library preparation involves extracting the RNA sequences from the cells in a sample of interest, conversion into stable cDNA, as well as multiple rounds of cDNA amplification to ensure that enough genetic material is available to perform next-generation sequencing [21]. The sequencing procedure then determines the exact nucleotide sequences contained in the sample, which are subsequently assigned to corresponding genes or taxa. For the contributions to this thesis, two HTS technologies are especially relevant.

(16S rRNA) Amplicon sequencing. 16S rRNA sequencing is a widespread sequencing method in microbiome analysis. It examines the highly variable regions of the 16S ribosomal RNA to determine the microbial composition of a sample (Figure 1.2A) [22]. 16S rRNA sequencing processes samples as a whole, pooling together the information about all microbes in the sample during library preparation. By matching the overlapping patterns of individual NGS reads, they can be clustered into operational taxonomic units (OTUs) or amplicon sequence variants (ASVs). Aligning the nucleotide patterns from OTUs or ASVs to known reference sequences then allows to interpret them as biological taxa. Finally, the abundance of each OTU/ASV/taxon in every sample is determined and aggregated into a count matrix with n samples and p OTUs/ASVs/taxa (Figure 1.2A) [8]. This matrix corresponds to $\tilde{\boldsymbol{X}}$ from Figure 1.1 and is compositional due to limitations in the number of cells, or library size, that can be sequenced in each sample [12].

Single-cell RNA sequencing (scRNA-seq). Contrary to microbial analysis, where each sample contains many different species with individual genomes, the analysis of human or animal tissues must rely on functional differences to discern between cell types. This can be achieved by comparing the mRNA expression patterns of individual cells through single-cell RNA sequencing (scRNA-seq) [9], giving a detailed picture of the protein synthesis processes inside each cell at the time of analysis. The main biotechnological difference between scRNA-seq and amplicon sequencing is an initial cell isolation step, where individual cells are separated from each other, usually through microfluidics, and tagged with a unique identifier (Figure 1.2B) [23, 24]. After mRNA extraction, transcription, amplification, and sequencing, each read can therefore be attributed to a gene and cell, leading to a matrix that shows the expression of each gene in each cell for a single sample. Although this gene expression matrix constitutes another instance of high-dimensional sequencing count data, compositional analysis is generally not seen as necessary here, as the data details the total mRNA contained in each cell instead of a representative subset. The idea of using compositional statistics for the analysis of gene expression data has however been contemplated recently [13].
To determine the function of each cell (e.g. stem cells or types of immune cells) in scRNA-seq, gene expression data requires careful statistical processing (Figure 1.2C) [5, 17, 25]. After filtering out low-quality reads and other sequencing artifacts, the data needs to be scaled and variance-stabilized to make reads from individual cells comparable.

Extracting the relevant information by low-dimensional embedding steps such as singular value decompositions and UMAP embeddings [26], followed by clustering, can determine clusters of cells with similar gene expression, which are commonly referred to as cell types (Figure 1.2C). Analysis of characteristic gene expression patterns can finally elucidate the biological function of each cell type. By aggregating the cells in each sample by their cell type, a count matrix with n samples and p cell types, corresponding to $\tilde{\boldsymbol{X}}$ in Figure 1.1, can be obtained. Again, this data is compositional due to the limited number of cells that can be processed from each sample [1].

Besides their compositionality, HTS abundance datasets possess some other characteristics that make statistical modeling challenging [5, 12, 17].

- **Dimensionality.** The biggest difference between amplicon and scRNA-seq data is in their dimensionality. While amplicon sequencing produces high-dimensional count matrices with hundreds or even thousands of features, scRNA-seq data usually only consists of up to 50 cell types. Nevertheless, both data types often contain more features than samples, requiring model selection processes to avoid underdetermined solutions.
- **Zero entries and overdispersion.** Since the logarithm is undefined for zero values, they must be replaced in the count matrix before applying logarithmic transformations. Furthermore, amplicon sequencing data often contains disproportional amounts of zero entries and overdispersed nonzero counts, necessitating the removal of rare taxa, aggregation to a higher taxonomic rank, or specialized models to combat these characteristics.
- **Feature associations.** Different types of bacteria or cells interact with each other, forming relationships that introduce correlation patterns beyond the compositional constraint. These patterns should be respected in a generative model.
- **Feature hierarchies.** The microbial taxa or cell types can be grouped hierarchically according to their taxonomy, phylogeny, or cell lineage. Depending on the level of aggregation, different insights can be gained from the data, ranging from more general descriptions of the high-level feature composition to fine-grained analyses of highly specific cell types or taxa.

Bacterial scRNA-seq. Very recent advancements aim to analyze within-species functional heterogeneity of bacteria through scRNA-seq technologies [10, 27, 28]. Due to the smaller size and more delicate structure of bacteria, as well as the lower concentration of bacterial mRNA, protocols for bacterial scRNA-seq alter the processes used for library preparation and sequencing [29, 30, 31, 32, 33, 34]. This also requires potential adaptations in the scRNA-seq data processing pipeline (Figure 1.2C).

1.2. Aims of this Work

Respecting compositionality when analyzing HTS data of the forms outlined above has proven to be a necessity rather than an optional step [12, 35, 36]. My works presented in

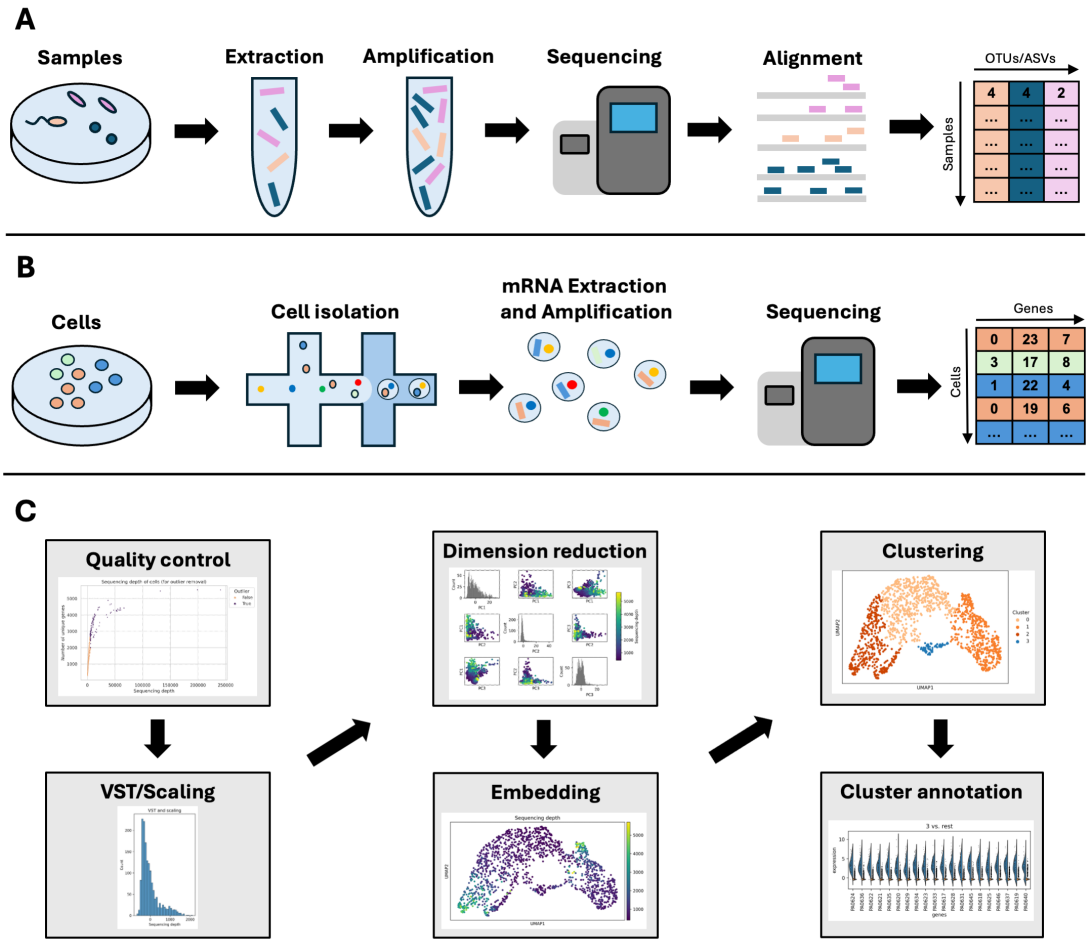


Figure 1.2.: **Experimental and data processing steps in high-throughput sequencing experiments. A)** Schematic representation of 16S rRNA amplicon sequencing. **B)** Schematic representation of single-cell RNA sequencing (scRNA-seq). **C)** Essential steps in scRNA-seq data analysis. *This figure is partially adapted from Figure 1 of contribution [4]*

this thesis mainly cover the development of generative models and methods for differential abundance testing for compositional data in light of the characteristics of HTS data described in the previous section. Hereby, I specifically focus on these challenges:

- How to design approaches that are valid for general low- to moderate-dimensional HTS data and not restricted to a specific technology or data type.
- How to model the impact of sample-specific metadata on the composition, select significant effects, and use this approach for differential abundance testing.
- How to include more complex settings, such as feature interactions or hierarchical ordering of the features.

Contributions [1] (Appendix A.1), [2] (Appendix A.2), and [3] (Appendix B) each approach these tasks with different statistical techniques and focus on specific data sources or characteristics.
A secondary objective of my work is the development of best practices and software pipelines for proper analysis of HTS data, making these tasks more accessible for researchers without extensive experience and detailed knowledge of the underlying statistical theory. Contribution [4] (Appendix C.1) is concerned with this exact task, detailing a pipeline that automatically determines suitable methods and hyperparameters for the analysis of bacterial scRNA-seq data. Furthermore, contribution [5] (Appendix C.2) discusses best practices for the analysis of general scRNA-seq data.

1.3. Outline

The following chapters serve as a general overview over the models and techniques used in the individual manuscripts. To point out the applicability to all kinds of HTS data, I will largely omit references to concrete technologies or biological connections, except when highlighting specific results from the contributions [1, 2, 3, 4, 5]. Instead, compositional HTS data will be presented as a generic matrix $\boldsymbol{X} \in \Delta_{p-1}^{n}$ with n samples and p features. In practice, these features should be interpreted as cell types, microbial taxa, OTUs/ASVs or similar, depending on the specific data at hand.
In Chapter 2, I will introduce various distributional families for modeling compositional data and develop an overarching framework that unites these approaches. In addition, I will shed light on the relationships between parameters in the particular distributions and their interpretation. Chapter 3 contains an overview over methods for parameter estimation used throughout the contributions. In Chapter 4, I will focus on modeling covariate data, selection of relevant model parameters, and differential abundance testing in the context of different estimation methods. I will discuss several other topics regarding modeling compositional data in Chapter 5, including approaches to handle count data, zero entries, and hierarchically ordered features. Chapter 6 introduces advanced techniques for data decomposition and generation, facilitating automatic determination of hyperparameters and model selection. I will give a short summary of the core results presented in the manuscripts in Chapter 7, and discuss future ideas beyond the scope of this thesis in Chapter 8.
The contributing manuscripts are included as appendices. In the first contribution ([1] in Appendix A.1), I introduce the `scCODA` model, a hierarchical Bayesian approach to generative modeling and differential abundance testing for compositional count data, specifically from scRNA-seq experiments. The second contribution ([2] in Appendix A.2) presents `tascCODA`, an extension of `scCODA` that allows for tree-aggregated differential abundance testing in general high-throughput sequencing datasets. Appendix B is dedicated to manus-cript [3], in which I introduce the `cosmoDA` model. This model uses a different approach to compositional generative modeling and DA testing, allowing the estimation of feature-feature associations, as well as removing the need for zero replacement. In contribution [4] (Appendix C.1), I present `BacSC`, a computational pipeline for automated statistical processing of gene expression data from bacterial single-cell RNA sequencing experiments. The final contribution ([5] in Appendix C.2) gives best practice

recommendations for the analysis of scRNA-seq data, where I contributed a section and an interactive tutorial resource on DA testing with `scCODA` and `tascCODA`.

1.4. Notation

These notations will be used throughout chapters 2-6. The notations in the contributions (Appendices A-C) might differ. Scalar quantities will be denoted in regular font, vectors in lowercase bold, and matrices with capital bold letters. The i-th element of vector $\boldsymbol{a}$ is denoted as a_i; $\boldsymbol{a}_{-i}$ is obtained by removing the i-th element from $\boldsymbol{a}$. The i-th row of a matrix $\boldsymbol{A}$ is denoted as $\boldsymbol{A}_{i,}$, the j-th column as $\boldsymbol{A}_{,j}$, and its element in position (i, j) as $A_{i,j}$. Removing the i-th row and j-th column from $\boldsymbol{A}$ results in $\boldsymbol{A}_{-i,-j}$. Power operations on vectors or matrices ($\boldsymbol{a}^n$ or $\boldsymbol{A}^n$) as well as multiplications with a scalar are carried out element-wise.

2. Distributions for Compositional Data

This section is an extension of the ideas presented in contribution [3].

Due to their constraint to the $(p-1)$-dimensional simplex and resulting dependence between features, HTS compositions require appropriate generative models that respect these properties [12, 15]. While the description of spurious correlations between measurements related to the same reference quantity dates back to Pearson [37], the connection of this phenomenon to proportional data was only made by Mosimann more than 60 years later [16]. Aitchison, who is often referred to as the "father" of compositional data analysis, provided the first formal description of the simplex as a vector space [15] and defined many fundamental concepts and principles for compositional statistics [38].

2.1. Parametric Distribution Families on the Simplex

Over the years, a multitude of parametric distributions with support on the simplex have been proposed [15, 16, 39, 40, 41, 42]. The following sections will introduce multiple traditional and more recent approaches (Figure 2.1) that follow Aitchison's ideas, albeit not always rigorously. To this end, consider a compositional dataset $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ with n samples and p features. Because of the compositionality constraint, $0 \preceq \boldsymbol{X} \preceq 1$ and $\sum_{j=1}^{p} X_{i,j} = 1 \; \forall i = 1, \ldots, n$. Let $\boldsymbol{x} = \boldsymbol{X}_{i,}$ for some $i \leq n$ be an arbitrary sample in the dataset.

2.1.1. The Dirichlet Distribution

The most straightforward distribution on the simplex is the Dirichlet distribution with probability density

$$p_{\mathcal{D}}(\boldsymbol{x}|\beta) \propto \prod_{j=1}^{p} x_j^{\beta_j - 1}, \tag{2.1}$$

where the concentration vector $\boldsymbol{\beta} \succ \mathbf{0}_p$ is the only parameter and defines the shape of the distribution. Its simple expectation $\mathbb{E}(\boldsymbol{x}) = \frac{\boldsymbol{\beta}}{\sum_j \beta_j}$ makes the Dirichlet distribution particularly easy to interpret. Furthermore, the Dirichlet distribution is the conjugate prior to the multinomial distribution (see Section 5.1), making it a popular choice for modeling compositional count data [1, 2, 43, 44, 45]. On the other hand, its simplicity also severely limits the flexibility of the Dirichlet distribution, as its covariance structure is always symmetric around the mode [39].

2.1.2. The Logistic Normal Distribution and Aitchison's A^{p-1} Family

To combat the inflexibility of the Dirichlet distribution, Aitchison and Shen [46] proposed the Logistic Normal distribution as an alternative to the Dirichlet. Define the additive logratio (ALR) transformation with reference component x_p as:

$$ALR(\boldsymbol{x}) = \log\left(\frac{\boldsymbol{x}_{-p}}{x_p}\right). \tag{2.2}$$

The Logistic Normal distribution is then constructed as a multivariate normal distribution over the ALR-transformed data:

$$p_{\mathcal{L}}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \propto \frac{1}{\prod_{j=1}^{p} x_j} \exp\left(-\frac{1}{2}\left(\log\left(\frac{\boldsymbol{x}_{-p}}{x_p}\right)-\boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\log\left(\frac{\boldsymbol{x}_{-p}}{x_p}\right)-\boldsymbol{\mu}\right)\right), \tag{2.3}$$

with $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T \in \mathbb{R}^{(p-1)\times(p-1)}$ positive semidefinite. This class of distributions allows for a flexible covariance structure due to the inclusion of $\boldsymbol{\Sigma}$, the covariance matrix of the ALR-transformed data, while the location vector $\boldsymbol{\mu}$ is related to the location vector of the Dirichlet distribution through $\boldsymbol{\beta}_{-p} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.
Aitchison [39] later combined the Dirichlet and Logistic Normal distributions into one joint distributional family by observing the similarity between their log-probability densities:

$$\log p_{\mathcal{D}}(\boldsymbol{x}|\boldsymbol{\beta}) \propto \sum_{j=1}^{p}(\beta_j - 1)\log(x_j)$$
$$\log p_{\mathcal{L}}(\boldsymbol{x}|\boldsymbol{\gamma},\boldsymbol{\beta}) \propto \sum_{j=1}^{p}(\beta_j - 1)\log(x_j) - \frac{1}{2}\sum_{\substack{j,k=1\\ j\neq k}}^{p}\gamma_{j,k}(\log x_j - \log x_k)^2.$$

The $A^{(p-1)}$ distribution is then defined equivalent to the log-density of the Logistic Normal distribution

$$p_{\mathcal{A}}(\boldsymbol{x}|\boldsymbol{\gamma},\boldsymbol{\beta}) \propto \sum_{j=1}^{p}(\beta_j - 1)\log(x_j) - \frac{1}{2}\sum_{\substack{j,k=1\\ j\neq k}}^{p}\gamma_{j,k}(\log x_j - \log x_k)^2, \tag{2.4}$$

where the density is proper if either $\boldsymbol{\gamma}$ is positive definite and $\boldsymbol{\beta} \succeq \mathbf{0}_p$, or $\boldsymbol{\gamma}$ is positive semidefinite and $\boldsymbol{\beta} \succ \mathbf{0}_p$. It is immediately visible that the log-density of the Dirichlet distribution, is identical to the density of the $A^{(p-1)}$ distribution with $\boldsymbol{\gamma} = 0$. In case of the Logistic Normal distribution, correspondence between the parameter sets $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ and $(\boldsymbol{\beta},\boldsymbol{\gamma})$ is shown in Table 2.1. One drawback of the logistic normal and $A^{(p-1)}$ distributions comes from the fact that the logarithm and ALR transformation are not defined if $\boldsymbol{x}$ contains zero entries. Therefore, these distributions do not cover the boundary of the simplex, and require replacement of zero entries before use.

2.1.3. Polynomially Tilted Pairwise Interaction Models

A more recent effort to combine the Dirichlet and Logistic Normal distributions was made by Scealy and Wood [40]. Their class of polynomially tilted pairwise interaction (PPI) distributions tries to unite the flexibility of the logistic normal distribution with the boundary behavior of the Dirichlet distribution:

$$p_{\mathcal{I}}(\boldsymbol{x}|\boldsymbol{A}^*, \boldsymbol{c}) \propto \prod_{j=1}^{p} x_j^{c_j - 1} \exp(\boldsymbol{x}^T \boldsymbol{A}^* \boldsymbol{x}), \tag{2.5}$$

with $\boldsymbol{A}^* \in \mathbb{R}^{p \times p}$ symmetric and $\boldsymbol{c} \succ -\boldsymbol{1}_p$. Explicitly stating the compositional constraint as $x_p = 1 - \sum_{j=1}^{(p-1)} x_j$ gives an alternative form with $p-1$ dimensions in the quadratic part [47]:

$$p_{\mathcal{I}}(\boldsymbol{x}|\boldsymbol{A}_L, \boldsymbol{b}_L, \boldsymbol{c}) \propto \prod_{j=1}^{p} x_j^{c_j - 1} \exp(\boldsymbol{x}_{-p}^T \boldsymbol{A}_L \boldsymbol{x}_{-p} + \boldsymbol{b}_L^T \boldsymbol{x}_{-p}). \tag{2.6}$$

Again, $\boldsymbol{A}_L \in \mathbb{R}^{(p-1)\times(p-1)}$ must be quadratic. The two parametrizations can easily be transformed into each other by splitting off the last row and column of $\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_L^* & \boldsymbol{A}_p^* \\ \boldsymbol{A}_p^{*T} & \boldsymbol{A}_{pp}^* \end{pmatrix}$. Then, $\boldsymbol{A}_{Li,j} = \boldsymbol{A}^*{}_{i,j} - 2\boldsymbol{A}_{p_i}^* + \boldsymbol{A}_{pp}^*$ and $\boldsymbol{b}_{Li} = 2(\boldsymbol{A}_{p_i}^* - \boldsymbol{A}_{pp}^*)$. Since $\boldsymbol{A}^*$ has one additional parameter, assume $\boldsymbol{A}_{pp}^* = 0$ for the reverse transformation. Then, $\boldsymbol{A}_{p_i}^* = \frac{1}{2}\boldsymbol{c}_i$, and $\boldsymbol{A}_{Li,j}^* = \boldsymbol{A}_{Li,j} + \boldsymbol{b}_{Li}$.

2.1.4. The Maximum Entropy Distribution

Another idea of defining a probability density on the simplex comes from the maximum entropy principle [48]. Maximizing the entropy of a distribution while matching the first and second moments of the data leads to the following formulation [41]:

$$\begin{aligned} p_{\mathcal{E}}(\boldsymbol{x}|\boldsymbol{M}, \boldsymbol{h}) &\propto \exp\left(\sum_{j=1}^{p} \left(h_j + \sum_{\substack{k=1 \\ k\neq j}}^{p} M_{j,k} x_k\right) x_j\right) \\ &= \exp(\boldsymbol{x}^T \boldsymbol{M} \boldsymbol{x} + \boldsymbol{h}^T \boldsymbol{x}), \end{aligned} \tag{2.7}$$

where $\boldsymbol{M} \in \mathbb{R}^{p\times p}$ is symmetric and has a zero diagonal and $h_p = 0$. The maximum entropy (ME) distribution differs from the PPI models only in the linear part, while the quadratic terms are identical, except for the constraints on $\boldsymbol{M}$ and $\boldsymbol{A}^*$. Furthermore, the ME distribution also accommodates zero entries in the data without requiring replacement.

2.1.5. a-b Power Interaction Models

The most general class of distributions on the simplex to date is the class of a-b power interaction models (PIM), introduced by Yu et al. [42]. Their formulation adds two

hyperparameters $a, b \geq 0$ to include different power transformations in the linear and quadratic parts:

$$p_{\mathcal{P}}(\boldsymbol{x}|\boldsymbol{K}, \boldsymbol{\eta}, a, b) \propto \exp(\frac{1}{2a}\boldsymbol{x}^{aT}\boldsymbol{K}\boldsymbol{x}^{a} + \frac{1}{b}\boldsymbol{\eta}^{T}\boldsymbol{x}^{b}), \tag{2.8}$$

with $\boldsymbol{K} \in \mathbb{R}^{p\times p}$ symmetric and an optional zero-sum constraint, $\boldsymbol{K}\mathbf{1}_p = \mathbf{0}_p$. If $a = 0$, define $\boldsymbol{x}^a$ as $\log(\boldsymbol{x})$ and replace $\frac{1}{2a}$ with 1, as the logarithm gives a smooth continuation of $\frac{1}{a}\boldsymbol{x}^a$, except for a shift independent of $\boldsymbol{x}$. Analogously, if $b = 0$, set $\boldsymbol{x}^b$ to $\log(\boldsymbol{x})$ and replace $\frac{1}{b}$ with 1. This makes the PIM the most flexible class of distributions shown here. If $a > 0$ and $b > 0$, PIM distributions also do not require zero replacement.
Note that the density defined for the PIM is always proper if $a > 0$ and $b > 0$. If $b = 0$, $\boldsymbol{\eta} \succ -1$ is required for equation 2.8 to form a proper density. If $a = 0$ and $b = 0$, $\log(\boldsymbol{x})^T\boldsymbol{K}\log(\boldsymbol{x}) \prec 0\ \forall \boldsymbol{x} \in \Delta_{p-1}$ is necessary [42].

2.2. Connections Between Distributions

Comparing the compositional distribution classes collected above, it is apparent that their densities all follow the same general quadratic form. All distribution classes have a location vector ($\boldsymbol{\mu}$, $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, $\boldsymbol{h}$, or $\boldsymbol{c}$) and - except for the Dirichlet distribution - an interaction matrix ($\boldsymbol{\Sigma}$, $\boldsymbol{\Gamma}$, $\boldsymbol{K}$, $\boldsymbol{M}$, or $\boldsymbol{A}^*$) as parameters. In fact, a-b power interaction models are not only the most flexible family, they actually form an overarching framework that encompasses the Dirichlet, Logistic Normal, $A^{(p-1)}$, Polynomially Tilted Pairwise Interaction (PPI), and Maximum Entropy (ME) distributions. To achieve equivalence between the a-b power interaction model (PIM) and the other distributional classes, it is necessary to fix the powers a and b and apply some constraints on the parameters $\boldsymbol{K}$ and $\boldsymbol{\eta}$ of the PIM. An overview over the relationship between all distributions as special cases of the PIM is given in Figure 2.1. Transformations between parameters and constraints necessary to achieve equivalence between the distributional classes are shown in Table 2.1.
The Dirichlet and Logistic Normal distributions are connected to the PIM through the $A^{(p-1)}$ class. Correspondences between $A^{(p-1)}$ models and the Dirichlet and Logistic Normal distributions were described in [39]. Noting that the $A^{(p-1)}$ class is derived from the log-density of the other two distributions, all three classes can be transformed into each other. Almost all of these transformations do not introduce additional restrictions on the parameters, except for a zero-sum constraint on the location parameter in case of a transformation from the $A^{(p-1)}$ to the Logistic Normal distribution. On the other hand, $A^{(p-1)}$ models are a special case of the PIM class with $a = 0$ and $b = 0$, corresponding to a logarithmic transformation of the data. To obtain equivalence between the classes, some restrictions to the PIM class must be made [42]. All elements of the location vector $\boldsymbol{\eta}$ are required to be larger than -1 to fall into the constraints of the Dirichlet distribution. Apart from the zero-sum constraint described above, $\boldsymbol{K}$ must also be positive definite (or positive semidefinite if $\boldsymbol{\eta} \succ -1$).
The ME model is obtained by setting $a = 1$ and $b = 1$ in the PIM class. Here, no further restrictions exist, as the density of both models is always proper.

The PPI models can be seen as a hybrid class between the $A^{(p-1)}$ and ME models, resulting from the PIM class through $a = 1$ and $b = 0$. Here, the constraint $\boldsymbol{\eta} \succ -1$ is necessary to achieve a proper distribution for the PIM class.
To simplify the notation in the upcoming chapters, the general form of the PIM (Eq. 2.8) with parameters $\boldsymbol{K}$ and $\boldsymbol{\eta}$ is used if not stated otherwise. The introduced techniques are applicable to all compositional models shown in this chapter, though.

2.3. Interpretation of Parameters

Despite their seemingly similar form, the parameters in the different distribution classes do not model the same quantities if the distribution-specific constraints on the parameters are not respected. While more investigation on the theoretical properties is necessary to achieve full interpretability of all distributions, there exist some interpretations of the parameters in individual classes:

Location. Only the Dirichlet distribution has an analytical solution for its mean, while the location parameter for the other distributions does not directly depict its expectation. Nevertheless, the location parameter in the PPI, $A^{(p-1)}$, Logistic Normal, and PIM ($b = 0$) classes shows some empirical similarity to the mean if feature associations are not too strong [3, 39, 40], with each component corresponding to one feature of the composition. While the quantitative interpretation is not fully clear, a qualitative assessment of significant changes in the location vector is thus possible.

Feature Associations. The interaction matrix in the Logistic Normal distribution gives the covariance matrix of the data after an additive logratio (ALR, Equation 2.2) transformation. While this quantity has no direct correspondence to the untransformed data, the covariance of the centered logratio (CLR) transformation approaches the empirical data covariance for large p [49]. Direct correspondence between the ALR and CLR covariance is achieved through a simple algebraic transformation [50]. In the case of $a = 0$ and if $\boldsymbol{K}$ is positive (semi-)definite, i.e. the distribution is proper, the transformations from Table 2.1 allow for an equivalent interpretation in the PIM class. For the other classes and if $a > 0$, the quantitative interpretation of the interaction matrix is unclear. Nevertheless, a qualitative assessment of sparse interaction matrices can detect significant associations between features [42].

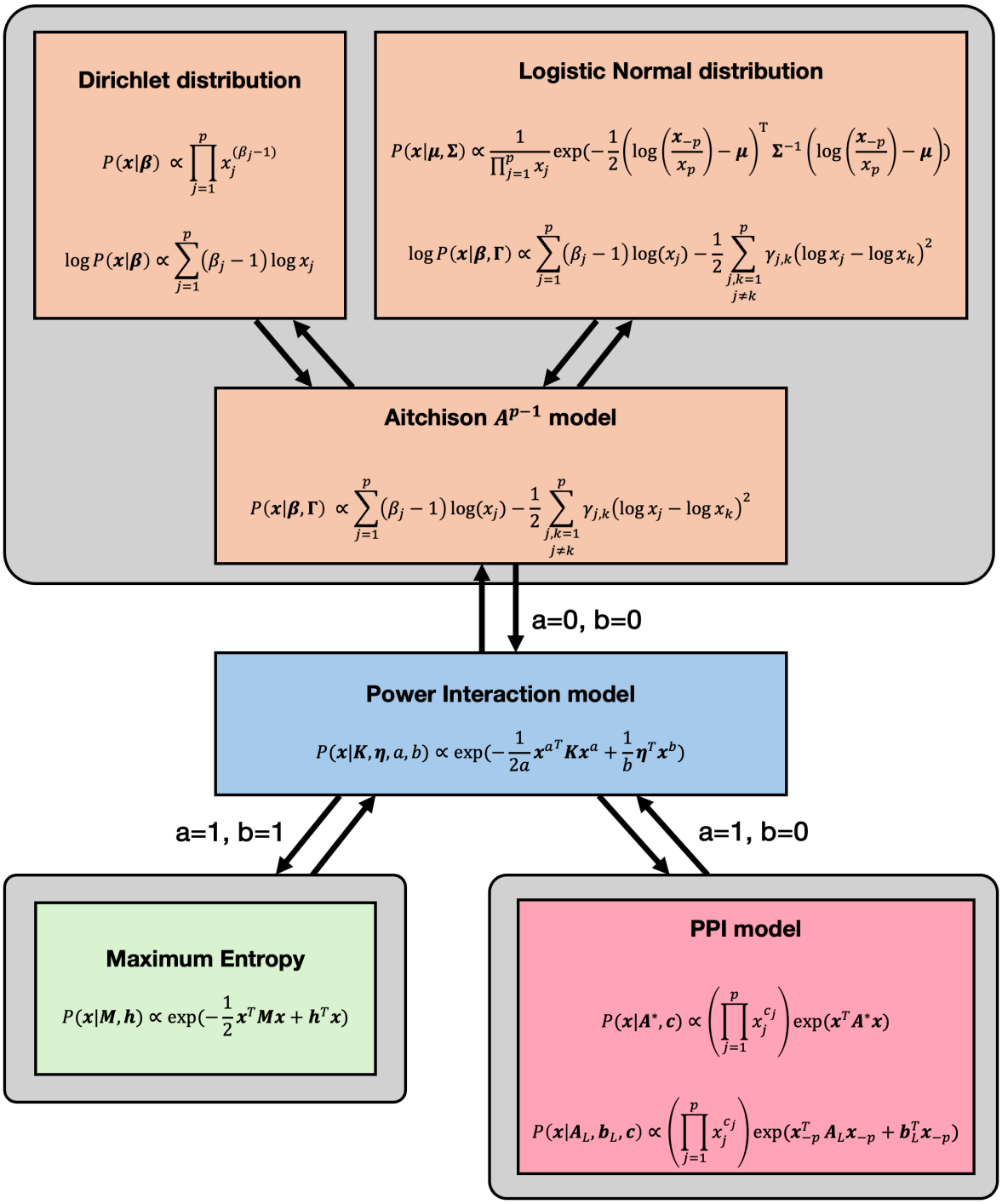


Figure 2.1.: **Overview over compositional distribution families and their probability densities.** Arrows show direct relationships achieved by reparametrization and setting the exponents in the a-b power interaction model to the values shown.

Distributions	**Transformations**	**Constraints**
LN $\to A^{(p-1)}$	$\gamma_{j,k} = -\frac{1}{2}\Sigma^{-1}_{j,k} \quad \gamma_{j,p} = \frac{1}{2}\sum_{k=1}^{p-1}\Sigma^{-1}_{j,k}$ $\boldsymbol{\beta}_{-p} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \qquad \boldsymbol{\beta}_p = -\sum_{j=1}^{p-1}\beta_j$	
$A^{(p-1)} \to$ LN	$\Sigma^{-1}_{j,k} = -2\gamma_{j,k} \quad \Sigma^{-1}_{j,j} = 2\sum_{\substack{j,k=1 \\ j\neq k}}^{p-1}\gamma_{j,k}$ $\boldsymbol{\mu}_{-p} = \boldsymbol{\Sigma}\boldsymbol{\beta}_{-p}$	$\boldsymbol{\beta}\mathbf{1}_p = 0$
PIM $\to A^{(p-1)}$	$\gamma_{j,k} = -\frac{1}{2}K_{j,k}$ $\boldsymbol{\beta} = \boldsymbol{\eta} + 1$	$a = 0;\ b = 0;\ \boldsymbol{\eta} \succeq -1;$ $\boldsymbol{K}\mathbf{1}_p = \mathbf{0}_p;\ \boldsymbol{x}^T\boldsymbol{K}\boldsymbol{x} > 0$
$A^{(p-1)} \to$ PIM	$K_{j,k} = -2\gamma_{j,k} \quad K_{j,j} = 2\sum_{\substack{j,k=1 \\ j\neq k}}^{p}\gamma_{j,k}$ $\boldsymbol{\eta} = \boldsymbol{\beta} - 1$	
PIM $\to$ PPI	$\boldsymbol{A}^* = \boldsymbol{K}$ $\boldsymbol{c} = \boldsymbol{\eta}$	$a = 1;\ b = 0;\ \boldsymbol{\eta} \succ -1$
PPI $\to$ PIM	$\boldsymbol{K} = \boldsymbol{A}^*$ $\boldsymbol{\eta} = \boldsymbol{c}$	$\boldsymbol{A}^*\mathbf{1}_p = \mathbf{0}_p$ (optional)
PIM $\to$ ME	$M_{j,k} = K_{j,k} \quad M_{j,j} = 0$ $\boldsymbol{h} = \boldsymbol{\eta}$	$a = 1;\ b = 1$
ME $\to$ PIM	$K_{j,k} = M_{j,k} \quad K_{j,j} = -\sum_{\substack{j,k=1 \\ j\neq k}}^{p} M_{j,k}$ $\boldsymbol{\eta} = \boldsymbol{h}$	

Table 2.1.: Correspondence between parameters and constraints of different compositional distribution families. *LN* denotes the Logistic Normal distribution, *PIM* the a-b power interaction models, *PPI* the polynomially tilted pairwise interaction models, and *ME* the maximum entropy distribution.

3. Parameter estimation methods for compositional distributions

The distribution families introduced in Chapter 2 allow to model compositional data with different degrees of flexibility. This chapter introduces parameter estimation methods to determine the specific values in the location vector and interaction matrix of these distributions that best describe a given compositional dataset $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, where $\boldsymbol{x} = \boldsymbol{X}_{i,}$ for some $i \leq n$. To this end, assume a general compositional distribution with probability density $p(\boldsymbol{x}|\boldsymbol{K}, \boldsymbol{\eta})$, location vector $\boldsymbol{\eta}$, and interaction matrix $\boldsymbol{K}$. This includes all distributions introduced in Chapter 2 through Table 2.1 (For a-b power interaction models, a and b are always fixed before estimation of $\boldsymbol{K}$ and $\boldsymbol{\eta}$). Define $\boldsymbol{\theta} = (\text{vec}(\boldsymbol{K}), \boldsymbol{\eta})$ as the vectorized collection of all entries in $\boldsymbol{K}$ and $\boldsymbol{\eta}$.
While the expressions in equations 2.1, 2.3, 2.4, 2.5, 2.7, and 2.8 all define proper probability densities (see Section 2.2 for potential conditions on the parameters), they were only stated proportionally to the actual densities though, missing the normalizing constant $1/\int_\Delta p(\boldsymbol{x}|\boldsymbol{\theta})\, dx$. Indeed, the normalizing constant only has an analytical solution for the Dirichlet, Logistic Normal, and $A^{(p-1)}$ models. Therefore, maximum likelihood estimation is only possible for these tractable distributions. For the other distribution families or more complex hierarchical models involving one of the compositional distributions, other strategies for parameter estimation are necessary. This chapter introduces two such techniques that I employed in `scCODA` [1], `tascCODA` [2], and `cosmoDA` [3]. `scCODA` and `tascCODA` in Appendix A both use Markov Chain Monte Carlo methods, while `cosmoDA` in Appendix B uses score matching optimization.

3.1. Markov Chain Monte Carlo Methods

Markov Chain Monte Carlo (MCMC) methods are a class of Bayesian inference algorithms that determine the parameters $\boldsymbol{\theta}$ of a probability density $p(\boldsymbol{X}|\boldsymbol{\theta})$ by repeatedly sampling from its unnormalized distribution. Bayes' theorem [51] states that the posterior distribution of parameters $p(\boldsymbol{\theta}|\boldsymbol{X})$ is proportional to the product of $p(\boldsymbol{X}|\boldsymbol{\theta})$ and the prior distribution $p(\boldsymbol{\theta})$ over the parameters:

$$p(\boldsymbol{\theta}|\boldsymbol{X}) \propto p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{3.1}$$

To determine the posterior distribution of $\boldsymbol{\theta}$, it is therefore sufficient to generate samples from $\boldsymbol{\theta}$ such that their distribution is equivalent to $f(\boldsymbol{\theta}) \equiv p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ [52]. MCMC methods achieve this goal by constructing a Markov chain $\Theta = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{t^*})$ of length t^* with stationary distribution $f(\boldsymbol{\theta})$ [53, 54].

The simplest way of constructing such a Markov chain is given by the Metropolis-Hastings algorithm [55, 56]. In every step $t = 1, \ldots, t^*$, the algorithm proposes a new state $\boldsymbol{\theta}'$ according to a proposal distribution $q(\boldsymbol{\theta}'|\boldsymbol{\theta}_{t-1})$ that depends on the previous state $\boldsymbol{\theta}_{t-1}$. The proposal is then accepted with probability

$$\alpha = \min\left(1, \frac{f(\boldsymbol{\theta}_t)q(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t)}{f(\boldsymbol{\theta}_{t-1})q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})}\right).$$

While the Metropolis-Hastings algorithm guarantees to sample from the posterior distribution in the limit, reaching this state may require a large number of iterations. Especially for high-dimensional problems with involved posterior distributions that have steep local maxima, the computational expense for reaching convergence of the Markov chain can be immense. To combat this issue, different variants of MCMC sampling exist that explore the posterior space more efficiently and achieve faster convergence as a result. Here, I briefly introduce two such approaches that I used in manuscripts `scCODA` ([1], see AppendixA.1) and `tascCODA` ([2], see AppendixA.2), respectively.

Hamiltonian Monte Carlo Sampling. Hamiltonian Monte Carlo (HMC) methods [57, 58] eliminate the choice of a proposal in the Metropolis-Hastings algorithm by using Hamiltonian dynamics to explore the relevant regions of the parameter space. Betancourt [59] gives a great introduction, which is summarized here.
In essence, HMC interprets the posterior space as a dynamic system by treating the states of the Markov chain as locations and introducing a set of latent momentum variables. The posterior distribution can therefore be interpreted as the potential energy landscape of this system. To generate a new proposal for the Metropolis-Hastings algorithm, HMC first randomly samples the momentum variables. Then, the state of the dynamic system after time Δt is determined through Hamilton's equations and used as the new proposal. Intuitively, the distance between the proposed HMC samples will therefore be close in regions with high posterior density and further apart when exploring a region of low posterior density. This process greatly reduces the autocorrelation between states when compared to a random walk and thus generally requires less sampling steps to obtain a good posterior sample.

No-U-turn Sampling. HMC is able to efficiently generate posterior samples, but requires tuning of the step size Δt to reach its full potential. In fact, the ideal step size depends on the geometry of the posterior distribution around the current state of the Markov chain. If Δt is too small, the process might move too slowly in each step, while a larger step size can cause oscillation around small high-density areas without exploring the maximum [59]. The No-U-Turn sampler (NUTS) [60] adaptively chooses a suitable Δt in each iteration. Therefore, exponentially increasing step sizes in both directions are simulated until a "U-turn", i.e. a reversal of the simulated trajectory's direction, is reached. Through this process, NUTS allows efficient exploration of the parameter space.

3.2. Score Matching Optimization

In contrast to MCMC methods that rely on costly simulations, score matching optimization is an analytical approach to parameter estimation of unnormalized distributions.

Given a family of parametric distributions $P_{\boldsymbol{\theta}}(\mathcal{D})$ defined on a domain $\mathcal{D} \subseteq \mathbb{R}^p$ with densities $p(\boldsymbol{x}|\boldsymbol{\theta})$ for each sample, score matching minimizes the Hyvärinen divergence [61] between p and the empirical data distribution $p_0(\boldsymbol{x})$:

$$J(\boldsymbol{\theta}) = \frac{1}{2} \int_{\mathcal{D}} p_0(\boldsymbol{x}) \; ||\psi(\boldsymbol{x}|\boldsymbol{\theta}) - \psi_0(\boldsymbol{x})||_2^2 \; dx. \tag{3.2}$$

Here, $\psi(\boldsymbol{x}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{\theta})$ and $\psi_0(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} p_0(\boldsymbol{x})$ are the Fisher score functions of p and p_0. Score matching is closely related to variational inference methods, which minimize the Kullback-Leibler (KL) divergence [62] $K(\boldsymbol{\theta}) = \int_{\mathcal{D}} p_0(\boldsymbol{x})(\log(p_0(\boldsymbol{x})) - \log(p(\boldsymbol{x}|\boldsymbol{\theta})))dx$ instead. By taking the squared difference between the gradients of the log-densities, the Hyvärinen divergence has one big advantage over the KL divergence: While the integral in the KL divergence depends on the normalizing constant and thus requires a tractable family of distributions, the normalizing constant in the Hyvärinen divergence vanishes through partial integration [61]:

$$J(\boldsymbol{\theta}) = \int_{\mathcal{D}} p_0(\boldsymbol{x}) \; \sum_{j=1}^{p} \left(\frac{\partial^2 \log(q(\boldsymbol{x}, \boldsymbol{\theta}))}{\partial x_j^2} + \frac{1}{2} \psi_j(\boldsymbol{x}, \boldsymbol{\theta})^2 \right) \; d\boldsymbol{x} + const, \tag{3.3}$$

where $q(\boldsymbol{x}|\boldsymbol{\theta})$ is the unnormalized density $p(\boldsymbol{x}|\boldsymbol{\theta})$. Under mild regularity conditions, minimizing the Hyvärinen divergence over all possible values of $\boldsymbol{\theta}$ provides a consistent estimator for the true parameters [61]. In practice, the integral in equation 3.3 is replaced with the mean over all samples to estimate $\boldsymbol{\theta}$. An especially convenient simplification of the optimization problem is possible for exponential-family-type models with densities $\log p(\boldsymbol{x}|\boldsymbol{\theta}) = \boldsymbol{\theta}^T a(\boldsymbol{x}) + b(\boldsymbol{x}) - c(\boldsymbol{\theta})$, where $a(\cdot)$ denotes the function for the sufficient statistics, $b(\cdot)$ the logarithm of the base measure, and $c(\cdot)$ the cumulant function. Here, the objective function reduces to a quadratic optimization problem [63]:

$$J(\boldsymbol{\theta}) = \frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\Gamma}(\boldsymbol{X}) \boldsymbol{\theta} + \boldsymbol{g}(\boldsymbol{X})^T \boldsymbol{\theta} + const, \tag{3.4}$$

where $\boldsymbol{\Gamma}(\boldsymbol{X})$ and $\boldsymbol{g}(\boldsymbol{X})$ are sample averages of functions in $\boldsymbol{X}$. This makes score matching optimization an appealing parameter estimation method for all compositional distributions from Section 2, as they are all exponential-family-type models.
However, the score matching estimator as described by Hyvärinen [61] is only valid for unbounded domains $\mathcal{D} = \mathbb{R}^p$. Generalizations of this formulation to the positive orthant [63, 64], oriented Riemannian manifolds [65], and general bounded domains [66] have since been made. The latter approach allows for score matching optimization of distributions constrained to the simplex and uses a weighting function $h(\boldsymbol{x}) = \left(\tilde{h} \circ \varphi \right)(\boldsymbol{x})$ to scale the score matching loss based on the (truncated) distance $\varphi(\boldsymbol{x})$ of $\boldsymbol{x}$ to the boundary of $\mathcal{D}$:

$$J(\boldsymbol{\theta}) = \frac{1}{2} \int_{\mathcal{D}} p_0(\boldsymbol{x}) \; ||\psi(\boldsymbol{x}, \boldsymbol{\theta}) \odot h(\boldsymbol{x})^{1/2} - \psi_0(\boldsymbol{x}) \odot h(\boldsymbol{x})^{1/2}||_2^2 \; dx. \tag{3.5}$$

Score matching for bounded domains was used for parameter estimation of Polynomially Tilted Pairwise Interaction models in [40, 47], and for a-b power interaction models (PIM) in [42]. For the `cosmoDA` model (Contribution [3], Appendix (Appendix B)), I extended the score matching on PIM distributions to include a covariate on the location vector.

4. Covariates and Differential Abundance Testing

Inferring the set of parameters ($\boldsymbol{K}$ and $\boldsymbol{\eta}$ or equivalent) in one of the compositional distributions from chapter 2 allows to describe a high-throughput sequencing dataset $\boldsymbol{X}$ with samples collected under the same conditions and circumstances. This approach is very useful to model the overall composition and associations between features in a homogeneous cellular population [40, 41, 42]. Many larger studies however do not only contain compositional abundance data from high-throughput sequencing, but consist of samples from different hosts or environments and according metadata information for each sample. The type of information depends on the biological context and can range from environmental (e.g. collection location, temperature, or pH value [67, 68]) and subject-specific covariates (e.g. age, sex, or dietary information [69, 70, 71]) to clinical information such as disease status or medical treatment information [72, 73, 74, 75]. Describing the impact of these covariates on the composition of cells or microbes can elucidate disease progression [72, 74], help in drug development [76], or explain evolutionary differences [77].
Particular interest lies not only in the quantification of compositional changes caused by a covariate, but also in separating biologically relevant effects from experimental noise. This chapter introduces a general concept for the description of compositional changes by incorporating the covariates into the distributions from Chapter 2, describes ways to select sparse parameter subsets in different modeling contexts, and gives an introduction to testing for statistical significance of these effects.

4.1. Handling Covariates in Compositional Models

The `scCODA` ([1], see Appendix A.1), `tascCODA` ([2], see Appendix A.2), and `cosmoDA` ([3], see Appendix B) models all follow the same conceptual design for covariate modeling. Let the matrix $\boldsymbol{Y} \in \mathbb{R}^{n \times d}$ denote the values of d covariates for each of the n samples. A linear model on the location vector can describe the effect of the covariates on the composition $\boldsymbol{X}$:

$$\boldsymbol{\eta} = \boldsymbol{\eta}_0 + \sum_{k=1}^{d} \boldsymbol{\eta}_k^T \boldsymbol{Y}_{i,k}. \tag{4.1}$$

For one sample $\boldsymbol{x} = \boldsymbol{X}_{i,}$ with $i \leq n$, this leads to a compositional model of the form

$$p_{\mathcal{P}}(\boldsymbol{x}|\boldsymbol{K}, \boldsymbol{\eta}, a, b) \propto \exp\left(\frac{1}{2a}\boldsymbol{x}^{aT}\boldsymbol{K}\boldsymbol{x}^a + \frac{1}{b}\left(\boldsymbol{\eta}_0 + \sum_{k=1}^{d} \boldsymbol{\eta}_k^T \boldsymbol{Y}_{i,k}\right)^T \boldsymbol{x}^b\right). \tag{4.2}$$

As stated in Section 2.3, the location vector is directly related to the mean of the composition for the Dirichlet, Logistic Normal, and $A^{(p-1)}$ distributions. For the a-b power interaction model (PIM), polynomially tilted pairwise interaction (PPI), and maximum entropy (ME) families, no direct association is possible, but the location vector still approximately represents the mean.

Modeling the influence of the covariates only on the location vector adopts the biological interpretation of a global interaction matrix that is not changed by environmental perturbations. This view is not able to characterize covariate-induced changes in the association pattern, but keeps the number of parameters at a reasonable level instead, even in settings where p is not small. In equation 4.1, each covariate introduces only p new parameters into the model, while an analogous model on the symmetric interaction matrix would lead to $(p(p+1))/2$ new parameters per covariate. Because the dimensions of most high-throughput sequencing datasets are not in the range of $n \gg p$ (see Section 1.1), the inclusion of covariates in the interaction matrix will often lead to misspecified models with more free parameters than samples.

Incorporating the covariate setup into the parameter estimation methods presented in Chapter 3 is straightforward. MCMC methods are particularly suited to accommodate all kinds of covariate setups through different prior structures in hierarchical Bayesian models [53, 78]. This in principle includes classes such as mixed or longitudinal models, although their discussion is beyond the scope of this thesis. For score matching, I developed an extension of the framework presented in [42] to include one covariate $\boldsymbol{y} \in \mathbb{R}^n$ in the form $\boldsymbol{\eta} = \boldsymbol{\eta}_0 + \boldsymbol{\eta}_1^T y_i$ in manuscript [3]. In theory, the ideas presented there also extend to multiple covariates, although this would require significant changes to the computational implementation of the score matching optimization.

4.2. Model Selection

Model selection in the context of this thesis refers to finding parameter sets for a compositional model that only have few nonzero entries, but still describe the data well. Selecting only a part of the parameters in a compositional model to be different from zero has multiple benefits. As seen in Section 1.1, cell type abundance datasets from scRNA-seq often have n in the same order of magnitude as p, while amplicon sequencing data typically even has many more features than samples. Inferring an interaction matrix with $(p(p+1))/2$ free parameters can therefore quickly lead to underdetermined models with more parameters than degrees of freedom [42]. Thus, limiting the number of free parameters is necessary to ensure model identifiability. Also, biologically relevant associations between species in microbial environments are believed to be sparse [79, 80]. Therefore, it is adequate to focus on the most relevant association effects, setting the rest of the entries in $\boldsymbol{K}$ to zero.

On the side of the location vector, enforcing sparsity on the covariate effects $\boldsymbol{\eta}_k;\ k \geq 1$ can select significant changes in the feature composition, providing a parsimonious and thus interpretable description. Depending on the chosen parameter estimation method, model selection can be performed with different techniques.

4.2.1. Penalized Model Estimation

For estimation methods that aim to minimize a loss function $L(\boldsymbol{X}, \boldsymbol{\theta})$ with respect to a dataset $\boldsymbol{X}$ and the parameters $\boldsymbol{\theta}$, for example the Hyvärinen divergence in the score matching estimator, $\mathcal{L}_1$ (LASSO) regularization [81] is the most common strategy for enforcing sparsity in the parameters. In its most basic form, the LASSO introduces a penalty based on the absolute values of all estimated parameters, which results in parameters being set to zero if their contribution to reduction of the loss does not outweigh the penalty:

$$L^*(\boldsymbol{X}, \boldsymbol{\theta}, \lambda) = L(\boldsymbol{X}, \boldsymbol{\theta}) + \lambda ||\boldsymbol{\theta}||_1. \tag{4.3}$$

The regularization strength $\lambda > 0$ must be tuned such that $\boldsymbol{\theta}$ is sparse and provides good out-of-sample predictive performance at the same time [82]. To achieve this goal, information criteria like the AIC [83], BIC, or eBIC [84, 85], or cross validation [86] are commonly employed. The information criteria all penalize the likelihood with the number of nonzero parameters and have the form

$$\begin{aligned} AIC(\boldsymbol{\theta}) &= 2\ \mathrm{Supp}(\boldsymbol{\theta}) - 2\log(L^*(\boldsymbol{X}, \boldsymbol{\theta}, \lambda)) \\ BIC(\boldsymbol{\theta}) &= \mathrm{Supp}(\boldsymbol{\theta})\log(n) - 2\log(L^*(\boldsymbol{X}, \boldsymbol{\theta}, \lambda)) \\ eBIC(\boldsymbol{\theta}, \gamma) &= \mathrm{Supp}(\boldsymbol{\theta})\log(n) - 2\log(L^*(\boldsymbol{X}, \boldsymbol{\theta}, \lambda)) + 2\gamma|\boldsymbol{\theta}|, \end{aligned}$$

where $|\boldsymbol{\theta}|$ is the number of entries in $\boldsymbol{\theta}$ and $\mathrm{Supp}(\boldsymbol{\theta})$ is its number of nonzero entries.
Cross validation tries to find a sparse solution that minimizes the prediction error on unseen data, thus finding a balance between bias and variance [87]. For this, the samples are first randomly split into k subsets $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ or folds of equal size. Iterating through each of the $i = 1, \ldots, k$ folds, define the test set as the i-th fold $\boldsymbol{X}_i$ and the training set as the union of all other folds, $\boldsymbol{X}_{-i}$. After obtaining the parameter set $\hat{\boldsymbol{\theta}}_i$ by minimizing $L^*(\boldsymbol{X}_{-i}, \boldsymbol{\theta}, \lambda)$, cross validation finds λ^* such that the cross validation error over all folds is minimized:

$$\lambda^* = \arg\min_{\lambda > 0} \sum_{i=1}^{k} L^*(\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}_i, \lambda). \tag{4.4}$$

In practice, it is often recommended to select models that are more sparse than the solution minimizing the cross validation error by selecting the largest λ within one standard error of the minimum [86].
For compositional distributions, not all parameters should be shrunk towards zero. In the interaction matrix $\boldsymbol{K}$, the diagonal elements are related to the variance of each feature and therefore not expected to be sparse. On the other hand, the off-diagonal pairwise feature associations $\boldsymbol{K}_{off}$ can be regularized to obtain a sparse and therefore biologically sound interaction matrix. A similar argument can be made for the location vector. Shrinkage on the intercept composition $\boldsymbol{\eta}_0$ will lead to inaccurate estimations of the mean composition. Performing model selection on covariate effects $\boldsymbol{\eta}_k$ can however uncover relevant perturbations of the composition under some conditions. A possible loss function for a compositional distribution could therefore be

$$L^*(\boldsymbol{X}, \boldsymbol{K}, \boldsymbol{\eta}, \lambda_1, \lambda_2) = L(\boldsymbol{X}, \boldsymbol{K}, \boldsymbol{\eta}) + \lambda_1 ||\boldsymbol{K}_{off}||_1 + \lambda_2 ||\boldsymbol{\eta}_1||_1. \tag{4.5}$$

Extending the regularized score matching estimator for PIM distributions [42, 88], I used regularized estimation for the PIM with covariates in the `cosmoDA` model ([3], see AppendixB), selecting the penalty strength λ_1 with cross validation.

4.2.2. Bayesian Spike-and-slab Priors

A Bayesian alternative to model selection are spike-and-slab priors which use a mixture of prior distributions to select between including and removing each variable instead of enforcing sparsity by penalizing the likelihood [89, 90]. These priors have the advantage that the hierarchical Bayesian formulation can easily accommodate prior assumptions, making them useful in settings with fewer samples, like single-cell experiments. Furthermore, the MCMC sampling procedure in a fully Bayesian approach allows to explore different combinations of selected variables rather than iteratively removing parameters from the model, which can improve model selection for highly correlated variables [91, 92].
The traditional formulation of the spike-and-slab prior, as introduced by Mitchell and Beauchamp [89], uses a Bernoulli prior τ_i to describe a parameter β_i as a mixture of two distributions, f_1 and f_2 (Figure 4.1A):

$$p(\beta_i|\tau_i) = \tau_i f_1 + (1 - \tau_i) f_2. \tag{4.6}$$

Here, f_1 is highly concentrated around 0 (the "spike") and corresponds to removing β_i from the model, while f_2 has a broader distribution (the "slab") and represents inclusion of β_i. In its most extreme case [89], the spike is chosen as the Dirac delta δ_0 at 0, and the slab is given by a uniform prior over the space of feasible values for β_i. Another common formulation represents both f_1 and f_2 as Gaussian distributions with low and high variance, respectively [90] (Figure 4.1A).

Hamiltonian Monte Carlo and Spike-and-slab Priors. Spike-and-slab priors use a discrete mixture of prior distributions to describe the parameter of interest. While this poses no problems for Metropolis or Gibbs samplers [93], Hamiltonian Monte Carlo methods require a fully continuous posterior distribution [59]. Therefore, the Bernoulli prior needs to be replaced with a continuous approximation when using HMC for parameter estimation. Two popular continuous approximations for the Bernoulli distribution in spike-and-slab priors are the Beta and Logistic Normal distribution [94] (Figure 4.1B). Spike-and-slab approaches for variable selection in compositional amplicon data were introduced by Wadsworth et al. [45]. In the `scCODA` model ([1], see AppendixA.1), I adjusted this approach to be suitable for scRNA-seq data and provided computational improvements through HMC sampling.

The Spike-and-slab LASSO Prior. *This section is adapted from the supplement of contribution [2].*
Another version of spike-and-slab priors was introduced by Ročková and George [95] and bridges the gap between the LASSO and Bayesian variable selection. This family of priors

uses a mixture of two double-exponential or Laplace distributions ψ_1 and ψ_0 to model the spike and slab portion:

$$\begin{aligned} p(\beta_i|\tau_i) &= \tau_i\psi_1(\beta_i) + (1-\tau_i)\psi_0(\beta_i) \\ \psi_1(\beta_i) &= \frac{\lambda_1}{2}e^{-\lambda_1|\beta_i|} \\ \psi_0(\beta_i) &= \frac{\lambda_0}{2}e^{-\lambda_0|\beta_i|}. \end{aligned} \tag{4.7}$$

Due to the nature of the double-exponential distribution, the mixture assumes a spike-and-slab form if $\lambda_0 \gg \lambda_1$ (Figure 4.1C). Conveniently, the spike-and-slab LASSO can be reformulated as a penalized likelihood function [95, 96] with penalty:

$$pen(\beta_i|\tau_i) = -\lambda_1|\beta_i| + \log(\frac{p^*_{\tau_i}(0)}{p^*_{\tau_i}(\beta_i)}), \tag{4.8}$$

where

$$p^*_{\tau_i}(a) = \frac{\tau_i\frac{\lambda_1}{2}e^{-\lambda_1|a|}}{\tau_i\frac{\lambda_1}{2}e^{-\lambda_1|a|} + (1-\tau_i)\frac{\lambda_0}{2}e^{-\lambda_0|a|}}. \tag{4.9}$$

If $\lambda_0 = \lambda_1$, $pen(\beta_i|\tau_i)$ reduces to $-\lambda_1|\beta_i|$, the LASSO penalty. In the `tascCODA` model ([2], see AppendixA.2), I used this property to achieve LASSO-like behavior of a spike-and-slab prior by setting λ_0 to a large constant value and increasing λ_1.

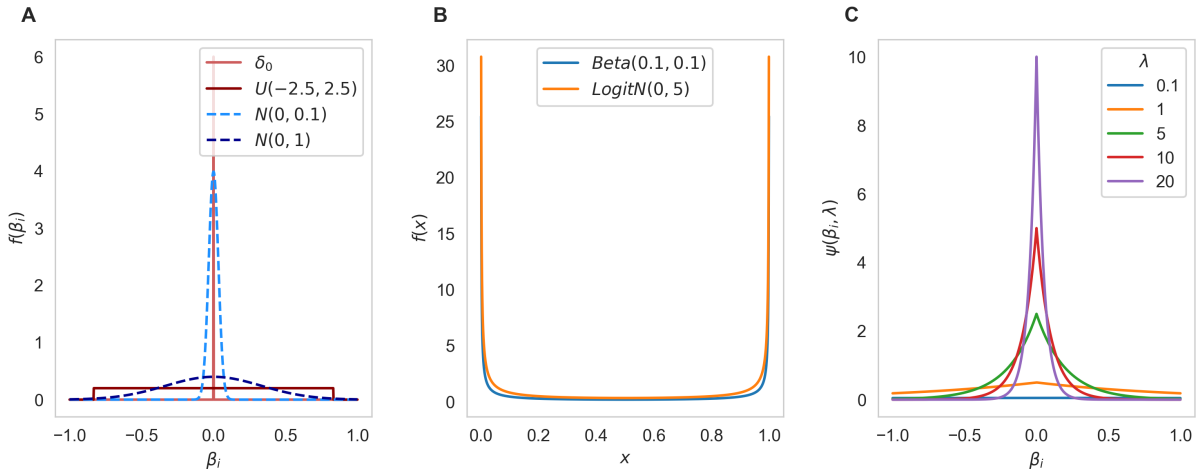


Figure 4.1.: **Spike-and-slab priors.** **A)** Components of the spike-and-slab priors by Mitchell and Beauchamp [89] (red, solid lines) and the normal mixture by George and McCulloch [90] (blue, dashed lines). **B)** The Beta and Logistic Normal distributions both approximate the Bernoulli distribution. **C)** The double-exponential prior for different values of λ.

4.3. Differential Abundance Testing

The central goal in modeling HTS data in light of external perturbations is to perform differential abundance (DA) testing, i.e. determining the set of covariate-feature associations that are different from zero with high confidence. During this task, the compositional nature of HTS abundance data is particularly relevant. Due to the natural negative correlation structure on the simplex, a shift in the abundance of one feature will cause a shift of all other feature abundances in the other direction. Ignoring this effect can easily lead to false-positive discoveries, grossly inflating the false discovery rate (FDR) [1, 97, 98]. Standard methods like the t-test or Wilcoxon rank-sum test are therefore infeasible for DA tests on HTS count data.
There is a multitude of DA testing methods that claim to provide FDR control for different data characteristics, modeling contexts, or sequencing technologies. Among these approaches, no universally "best" method can be determined, and benchmarking studies provide contradicting results [99, 100, 101, 102]. While a full description of the features and assumptions of all approaches is beyond the scope of this work, I want to briefly mention a few recurring ideas:

- ALR-like compositional approaches with reference features [1, 98, 103, 104]
- Correction of biases caused by the sampling procedure or outliers. This group includes the ANCOM family [105, 106, 107, 108], LinDA [109], as well as other methods [110, 111]
- Inclusion of feature associations, as proposed in [3, 112].
- Zero-inflated regression approaches [44, 113, 114] or avoiding zero replacement [3, 115, 116].
- Other ideas, such as differential neighborhoods [117] or tree-based hierarchical testing [2, 118].

At their core, most of these methods use regression-based approaches similar to Section 4.1, but differ in their modeling strategy and testing or variable selection procedure. Using one of the compositional models presented in Chapter 2 or transforming the data into unconstrained space, for example through a logratio transformation, ensures that the compositional constraint is respected.
For simplicity, assume that the binary group indicator is the only modeled covariate, i.e. $d = 1$ in equation 4.1. By testing the hypothesis

$$H_0 : \eta_{1j,k} = 0 \quad \text{vs.} \quad H_1 : \eta_{1j,k} \neq 0, \tag{4.10}$$

on every entry of $\boldsymbol{\eta}_1$ from equation 4.1, significant effects on individual features can be determined. For the test statistic, different types of studentized test statistics that relate the estimated mean of $\eta_{1j,k}$ to its estimated variance are commonly used [107, 109]. To fully control the FDR, the resulting p-values must finally be corrected for multiple testing [119, 120, 121]. Then, all parameters $\eta_{1j,k}$ with an adjusted p-value smaller than a FDR level α are differentially abundant. In the `cosmoDA` model ([3], see AppendixB), I used this

exact approach to perform DA testing in light of feature associations. Here, a consistent estimator for the variance of $\boldsymbol{\eta}_1$ comes directly from the score matching optimization [40].

Differential Abundance Testing with Spike-and-slab Priors. For approaches using MCMC methods and variable selection through spike-and-slab priors, hypothesis testing is not possible due to a lack of appropriate variance estimators. Instead, determination of credible effects and FDR control is possible through thresholding of the posterior inclusion probability [122, 123]. Given the parameter estimates $\beta_{il} = \tau_{il} f_{1l} + (1-\tau_{il}) f_{2l}$, $l = 1, \dots, K$ in a Markov chain of length K, the posterior inclusion probability (PIP) is the share of samples where the spike-and-slab prior is in the "slab" portion:

$$PPI(\beta_i) = \frac{|\{l : \tau_{il} = 0\}|}{K}. \tag{4.11}$$

A higher PIP therefore corresponds to a higher probability of β_i being nonzero. For the continuous approximations of a true Bernoulli spike-and-slab prior, one can calculate the PIP as the of samples where β_{il} is smaller than a certain value instead. For a fixed level α, a direct posterior probability approach as proposed in [124] can estimate the FDR:

$$FDR = \frac{\sum_{i=1}^{p} (1 - PPI(\beta_i)\ I(PPI(\beta_i) > \alpha))}{\sum_{i=1}^{p} I(PPI(\beta_i) > \alpha)}, \tag{4.12}$$

where I is the indicator function. Alternatively, a Beta prior on the prior probability of the selection variable τ_i leads to automatic FDR control without the need to threshold the PIP after the MCMC sampling [90]. In the spike-and-slab LASSO prior, this threshold is given as

$$\delta = \frac{1}{\lambda_0 - \lambda_1} \log(1/p^*_{\tau_i}(0) - 1), \tag{4.13}$$

where $p^*_{\tau_i}(0)$ as in equation 4.9 [95]. Only posterior parameter values exceeding this threshold are deemed as credibly different from 0 and thus the associated effects are differentially abundant. Because these Bayesian procedures do not produce p-values for the effects, these are referred to as credible instead of significant.
I implemented the direct posterior probability approach in `scCODA` ([1], see Appendix A.1), while I used the Beta prior with a practical significance threshold in `tascCODA` ([2], see Appendix A.2).

5. Further Challenges in Compositional Abundance Data Modeling

Chapters 2 - 4 provide an extensive toolbox to model compositional high-throughput sequencing data with and without covariates, perform efficient parameter estimation, and test for differential abundances. This framework allows to account for most of the characteristics of compositional HTS data outlined in Section 1.1, including underdetermined data with $p > n$ through model selection, overdispersion and feature associations through the inclusion of an interaction matrix, as well as covariates. This chapter discusses how to approach the remaining challenges from Section 1.1 - zero entries in the data and hierarchical ordering of the features - and outlines strategies to account for the fact that high-throughput sequencing data comes in the form of count data.

5.1. Handling of Count Data

The compositional modeling framework described so far is based on data located on the $(p-1)$-dimensional probability simplex, $\boldsymbol{X} \in \left(\Delta^{(p-1)}\right)^n$. However, HTS abundances occur in the form of count data, i.e. $\tilde{\boldsymbol{X}} \in \mathbb{N}_0^{(n \times p)}$. There are two natural approaches to bridge the gap between compositional count data and proportions defined on the simplex.
The straightforward solution simply uses the closure operation (Equation 1.2) to divide every sample in the data by its library size, i.e. the number of cells sequenced:

$$\boldsymbol{X}_{i,} = C(\tilde{\boldsymbol{X}}_{i,}) = \frac{\tilde{\boldsymbol{X}}_{i,}}{\sum_{j=1}^{p} \tilde{X}_{i,j}}. \tag{5.1}$$

The closure over $\tilde{\boldsymbol{X}}$ treats every sample purely as relative abundances and allows for a truly compositional view on the data. I used this approach for a-b power interaction models in combination with score matching optimization in `cosmoDA` ([3], see Appendix A.2). Similarly, gene expression data is usually scaled to a common sum by taking the closure and then multiplying the resulting proportions with a constant scaling factor s [5]. This embeds all samples in a scaled standard simplex $s\Delta_{p-1}$. I utilized this scaling in the context of bacterial single-cell sequencing in the `BacSC` pipeline ([4], see Appendix 6.3).
While the closure makes all samples comparable, it also completely eliminates any information about each sample's library size (or sequencing depth) from the data. This procedure is justified from a biological point of view, as the true absolute cell abundances of the analyzed populations cannot be recovered through sequencing and only relative information remains. When viewing high-throughput sequencing as a statistical process instead of a "black box" method for generating compositional data, the number of cells

in a sample does carry information about the sample uncertainty though. Assuming no biases in the sequencing process, a larger average library size can depict the true proportions in each sample more accurately. For very rare features, the sequencing depth can even significantly impact the probability of being detected at all. A multinomial sampling step on the proportions determined by the compositional model can account for varying sample uncertainty [125, 126]. The multinomial distribution takes a p-dimensional vector of proportions $\boldsymbol{\alpha} \in \Delta^{(p-1)}$ and draws from this distribution N times:

$$p_{\mathcal{M}}(\boldsymbol{x}|\boldsymbol{\alpha}, N) = \frac{N!}{\prod_{k=1}^{p} x_k!} \prod_{k=1}^{p} \alpha_k^{x_k}. \tag{5.2}$$

When modeling a concrete dataset, each sample's sequencing depth N_i can be directly determined from the data. Because the multinomial distribution is a conjugate prior to the Dirichlet distribution [16], the Dirichlet-Multinomial compound distribution is often used in hierarchical Bayesian models [45, 127]. The `scCODA` ([1], see Appendix A.1) and `tascCODA` ([2], see Appendix A.2) models utilize the Dirichlet-Multinomial distribution for modeling HTS count data.

5.2. Zero Entries

Virtually all HTS data - abundance as well as gene expression - contains some features that were not detected in all samples, leading to zero entries in $\boldsymbol{X}$. These can occur due to different reasons [106, 128]:

- True absence of a feature in a tissue or environment, also known as biological zeroes.
- Rare features that are present in the tissue or environment, but were not collected in the sampling process due to the limited depth of sampling. These are also known as sampling zeroes.
- Technical zeroes - features that were not detected for technical reasons, although they were present in the sample.

These zero entries pose a problem when applying logarithmic data transformations, as $\log(0)$ is undefined. Any transformation that includes taking the logarithm of entries in $\boldsymbol{X}$ thus requires replacement of zero entries. This includes all logratio transformations as well as every distribution from Chapter 2 that can be represented as an a-b power interaction model with either $a = 0$ or $b = 0$. While there exist many different approaches for zero imputation that often try to distinguish between these cases, their discussion is beyond the scope of this thesis. Extensive comparisons between zero replacement methods can e.g. be found in [128]. A simple, yet effective strategy replaces every zero entry in $\boldsymbol{X}$ with a constant, small value c, also known as a pseudocount. These pseudocounts are usually chosen to be either 1 or a smaller, positive value, for example 0.5. This strategy has generally proven to be a good approximation, although it does alter the measured composition, especially for rare features [128]. I used constant zero replacement with a value of 0.5 in `scCODA` ([1], see Appendix A.1), `tascCODA` ([2], see Appendix A.2), and `cosmoDA` ([3], see Appendix B).

An important relation between the value chosen for the pseudocount and the overdispersion assumed in the data is given by [129] in the context of variance-stabilizing transformations for scRNA-seq gene expression data. Consider a log-transformation after scaling the total number of reads in each cell to a constant value, e.g. the median sequencing depth over all cells:

$$\boldsymbol{Z}_{i,j} = \log(\frac{\tilde{\boldsymbol{Z}}_{i,j}}{m_i} + c_j), \tag{5.3}$$

where $m_i = \frac{\sum_{j=1}^{r} \tilde{\boldsymbol{Z}}_{i,j}}{\text{median}_{k=1}^{m}(\sum_{j=1}^{r} \tilde{\boldsymbol{Z}}_{k,j})}$ are the cell-specific scaling factors. If the data follows a Negative-Binomial distribution, choosing the pseudocount as $c_j = \frac{1}{4d_j}$, where d_j is the feature-specific overdispersion in the data, gives the best approximation to the variance stabilizing transformation determined through the delta method [130]. I employed the logarithmic transformation with feature-specific pseudocounts in `BacSC` ([4], see Appendix C.1 and Section 6.3).

The remainder of this section is based on contribution [3].
Power transformations of the form $\boldsymbol{X}^{\phi}$, $0 < \phi < 1$ are another data transformation strategy for compositional data. Because they are not undefined on the boundaries of the simplex, they completely eliminate the need for zero imputation. In fact, the Box-Cox transformation $\frac{1}{\phi}(x^{\phi}-1)$ [131] approaches the logarithm for decreasing values of the power ϕ (Figure 5.1):

$$\lim_{\phi \to 0} \frac{1}{\phi}(x^{\phi} - 1) = \log(x). \tag{5.4}$$

The power transformation used in a-b power interaction models (see Equation 2.8) with equal powers $a = b = \phi$ is similar to a Box-Cox transformation, albeit not equivalent, as it is missing a factor $\frac{1}{a}$ as well as the subtraction of 1. In `cosmoDA` ([3], see Appendix B), I introduced scaling factors in the score matching optimizer to approximate Box-Cox-like behavior for the estimated parameters $\boldsymbol{K}$ and $\boldsymbol{\eta}$ and compared how zero replacement and different power transformations influence differential abundance results in HTS.
Box-Cox-like transformations can also approximate logratio transformations. In particular, a scaled and closed version of the Box-Cox transformation converges to the centered logratio (CLR) transformation [132]:

$$\lim_{\phi \to 0} \frac{1}{\phi} \left(p \frac{\boldsymbol{x}^{\phi}}{\sum_{k=1}^{p} x_k^{\phi}} - 1 \right) = \log \left(\frac{\boldsymbol{x}}{g(\boldsymbol{x})} \right) \equiv CLR(\boldsymbol{x}), \tag{5.5}$$

where $g(\boldsymbol{x})$ is the geometric mean. This result is used by [133] to determine the value of ϕ that best preserves isometry to the zero-replaced and CLR-transformed data by maximizing the Procrustes correlation between the PCA embeddings of both matrices. The results in contribution [3] show that the same strategy can also be used to approximate the ALR transformation.

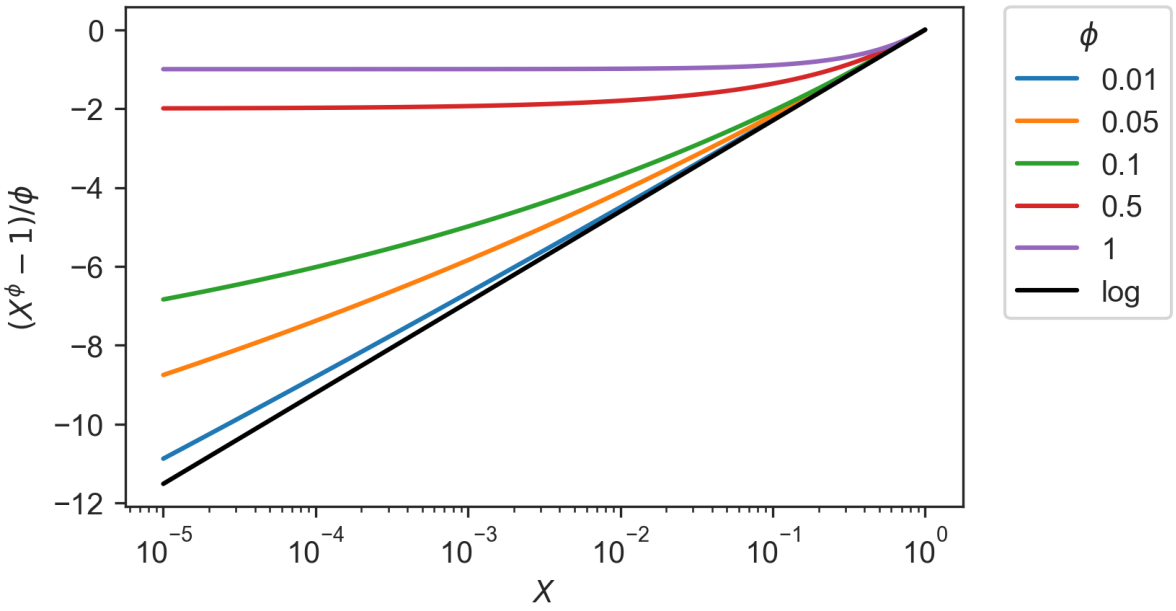


Figure 5.1.: **Relationship between the logarithm and Box-Cox transformations.** For decreasing exponents ϕ, the Box-Cox transformation converges to the logarithm. *This figure is part of Figure 1 in contribution [3].*

5.3. Tree Structures in HTS Data

This section is based on contribution [2].
Another HTS data characteristic that can be incorporated into generative modeling is an inherent hierarchical ordering of features. In both amplicon as well as scRNA-seq analysis, these orderings come in form of a tree that groups the features either based on expert biological knowledge or through hierarchical clustering algorithms. In microbiome analysis, these structures are based on taxonomy [134, 135], or phylogenetic similarity [136]. For scRNA-seq, cell lineage hierarchies or hierarchical clusterings based on gene expression patterns [137] can generate such tree structures.
These hierarchies can be used to aggregate features into more general groups for a broader analysis that is easy to interpret, or split them up for more nuanced and less parsimonious insights. Usually, the level of aggregation is fixed before the analysis and the respective feature groups are combined through summation. On the other hand, hierarchical feature selection methods [138, 139] are able to select a sparse set of important effects that is not confined to a predetermined aggregation level by introducing auxiliary variables for all aggregation levels of the tree and using adaptive penalized estimation. In the `tascCODA` model ([2], see Appendix A.2), I developed such a tree-based feature selection scheme for simultaneous DA testing on all aggregation levels. The approach uses a hierarchical Bayesian Dirichlet-Multinomial model with spike-and-slab LASSO priors from Section 4.2.2, employing adaptive penalization strengths for each node based on their position in the tree.

6. Statistically Valid Data Decomposition and Generation Schemes

This chapter is in large parts based on contribution [4].
Besides distributional parameters that are estimated to fit the data, e.g. through the methods introduced in Chapter 3, many statistical methods contain additional hyperparameters that need to be tuned. Common examples of such hyperparameters in high-throughput sequencing data analysis are the penalization strength of LASSO regularization (see Section 4.2.1), the latent dimensionality in truncated singular value decomposition [140], or the resolution parameter of a clustering algorithm [141, 142]. Hyperparameter selection aims to find a parameter value that fits the data at hand, without overfitting or producing false discoveries, therefore generalizing well to unseen data [143]. Two such approaches were already touched upon in Section 4.2.1, where cross validation and information criteria were presented as ways to select the regularization strength of the LASSO.
This chapter introduces data thinning [140], a data splitting technique that enables hyperparameter selection in unsupervised settings. Furthermore, approaches to generate valid null data for hyperparameter determination and false discovery rate control in hypothesis testing are discussed. The final section in this chapter describes the use of data thinning and null data generation within the `BacSC` pipeline (Contribution [4], Appendix C.1).

6.1. Data Thinning

To illustrate the idea of data thinning, also known as count splitting, it is helpful to showcase the inadequacy of traditional sample splitting procedures for unsupervised learning tasks first. For this, consider a dataset $\tilde{\boldsymbol{Z}} \in \mathbb{R}^{m \times r}$ with m samples and r features. In supervised learning, an additional ground truth $\boldsymbol{y} \in \mathbb{R}^m$ is given for each sample. Sample splitting, such as a single fold of cross validation, can be used to select the value of a hyperparameter λ, e.g. the penalization strength in LASSO regression (see Section 4.2.1), such that the associated solution $g^*_\lambda(\cdot)$ generalizes well to unseen data. For this, randomly divide the data into a training dataset $(\tilde{\boldsymbol{Z}}_{train}, \boldsymbol{y}_{train}) \in (\mathbb{R}^{m_1 \times r}, \mathbb{R}^{m_1})$ and a test dataset $(\tilde{\boldsymbol{Z}}_{test}, \boldsymbol{y}_{test}) \in (\mathbb{R}^{m_2 \times r}, \mathbb{R}^{m_2})$ with sample sizes $m_1 + m_2 = m$ (Figure 6.1A). For a fixed value of λ, the best solution $g^*_\lambda(\cdot)$ on the training dataset is first computed by optimizing a performance indicator (e.g. a loss function) $L(\boldsymbol{y}_{train}, g_\lambda(\tilde{\boldsymbol{Z}}_{train}))$ with respect to g_λ. The test error $L(\boldsymbol{y}_{test}, g^*_\lambda(\tilde{\boldsymbol{Z}}_{test}))$ (c.f. equation 4.4) then examines how well this solution fits the previously unseen test data. Minimizing this quantity with respect to λ yields a latent parameter with good generalization ability [54].

Sample splitting is only possible for supervised learning tasks though, as it requires a ground truth $\boldsymbol{y}$ to evaluate the performance indicator. In unsupervised hyperparameter selection settings, no ground truth $\boldsymbol{y}$ independent of $\tilde{\boldsymbol{Z}}$ exists, thus invalidating the procedure [144]. As an example, consider selecting the rank k of a truncated singular-value decomposition (SVD) $\hat{\boldsymbol{Z}}$ of $\tilde{\boldsymbol{Z}}$. To evaluate the quality of an SVD, it must be compared to the initial data, making $\tilde{\boldsymbol{Z}}$ the ground truth. This creates a scenario where the ground truth is also used to calculate the solution ($\hat{\boldsymbol{Z}}$), which would always lead to the maximal number of dimensions $k = \text{rank}(\tilde{\boldsymbol{Z}})$ being selected [145, 146].
For datasets with a convolution-closed data distribution, Neufeld et al. [140] propose count splitting as an alternative to sample splitting. A distribution $P(\theta)$ is convolution-closed in θ, if for two independent realizations $\tilde{\boldsymbol{Z}}_1 \sim P(\theta_1), \tilde{\boldsymbol{Z}}_2 \sim P(\theta_2)$, it holds

$$\tilde{\boldsymbol{Z}}_1 + \tilde{\boldsymbol{Z}}_2 \sim P(\theta_1 + \theta_2). \tag{6.1}$$

Many count distributions, such as the Poisson or negative binomial distribution, fulfill this condition with respect to their location parameters [147]. A convolution-closed distribution allows to split the data entry-wise such that $\tilde{Z}_{i,j} = \tilde{Z}_{i,j_{train}} + \tilde{Z}_{i,j_{test}} \quad \forall i, j$ and both the train and test dataset are part of the same distribution family as the full data [140] (Figure 6.1B). Because both $\tilde{\boldsymbol{Z}}_{train}$ and $\tilde{\boldsymbol{Z}}_{test}$ now follow the same distribution and contain information about the same samples, it is possible to find a solution (e.g. a truncated SVD) on the training data and directly evaluate it on the test data. Therefore, the test data is unseen before evaluation, which makes hyperparameter selection through minimization of the test error possible. For Poisson-distributed data, count splitting can be performed by simple Binomial allocation, while data with a Negative-Binomial distribution requires a Dirichlet-Multinomial sampling process [140, 146]. Following the same principles, it is also possible to create more than two data splits through count splitting, enabling multifold or leave-one-out approaches similar to cross validation [140].
I used count splitting for two tasks in the `BacSC` pipeline in contribution [4], determining the latent dimensionality k of the principal component embedding as well as the resolution parameter for Louvain or Leiden clustering [141, 142]. More details on these approaches follow in Section 6.3.

6.2. Randomized and Synthetic Null Data

Hyperparameter optimization methods commonly use metrics like loss functions or accuracy measures to determine the optimal value of λ by selecting the best average score over all data points. In some cases, for example when evaluating the quality of a UMAP embedding [26], the average performance over all data points is of secondary interest though. Instead, it is more important to minimize the number of data points that produce unwanted results, i.e. where the performance metric falls under a certain threshold value [148]. Often, such a threshold is not straightforward to define, since the range and distribution of values for a performance metric is usually not known beforehand. Instead, a data-driven approach can be chosen by using the performance metric's distribution on a negative control or null dataset without any signal as a baseline (Figure 6.2A). In this case, a data point produces an unwanted result if the value of its performance metric

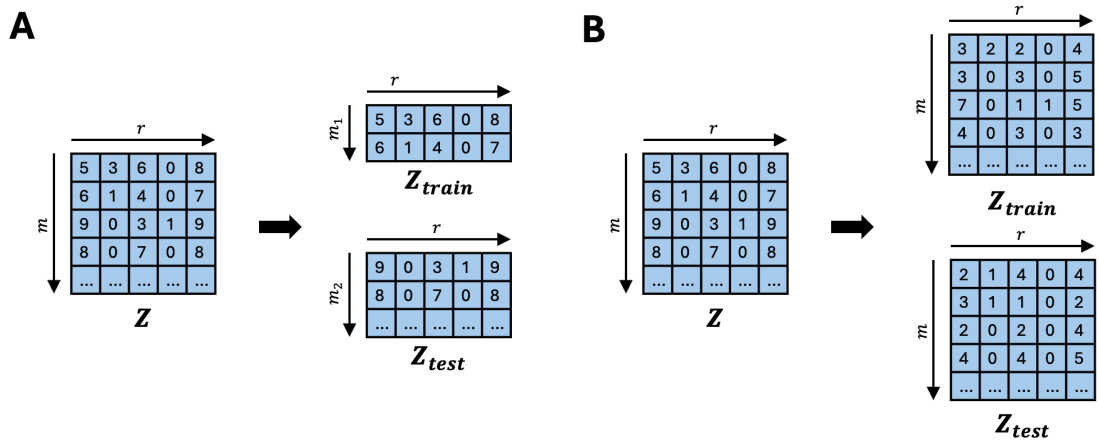


Figure 6.1.: **Count splitting. A)** Sample splitting divides the samples into train and test data. **B)** Count splitting partitions each data entry, thus achieving train and test data with comparable distributions and equal dimensionality.

lies below a certain quantile of the null distribution on the negative control data. Such negative control comparisons are closely tied to hypothesis tests, essentially evaluating the null hypothesis of no difference between the null data and the actual dataset [149].
A valid null dataset must have similar properties as the data of interest, but contains no signal that is relevant to the performance metric [150]. If no such null data was generated through a negative control experiment, several synthetic generation options exist. Analogous to permutation hypothesis tests [151], one strategy to obtain synthetic null data is permutation of the entries $\hat{\boldsymbol{Z}}_{1,j}, \ldots, \hat{\boldsymbol{Z}}_{m,j}$ for every feature $j = 1, \ldots, r$ in the dataset (Figure 6.2A). This strategy preserves the mean and variance of each feature, but removes any grouping or correlation structure, making it a valid null dataset for many tasks, including the evaluation of low-dimensional embeddings or clusterings [148].
In cases where the correlation patterns between features also need to be preserved, for example differential gene expression testing [152], other null data generation methods must be employed. One strategy for synthetic data generation that can preserve the mean, variance, and correlation patterns of arbitrary data are copula approaches [153] or the related Normal-to-anything method [154]. In amplicon sequencing, Normal-to-anything is used in the SPIEC-EASI pipeline [49] to generate synthetic data with specific correlation structures, while the scDesign family [155, 156, 157] provides copula approaches for generating synthetic gene expression data from scRNA-seq experiments.
At their core, all copula-based data generation methods follow the same principle (Figure 6.2B). First, parametric distributions $P_j(\cdot, \boldsymbol{\theta}_j)$ with parameters $\boldsymbol{\theta}_j$ are fit to the marginal distribution of each feature $j = 1, \ldots, r$ (Genes in Figure 6.2B) in the dataset $\tilde{\boldsymbol{Z}} \in \mathbb{R}^{m \times r}$. Here, the marginal distributions $P_j(\cdot, \boldsymbol{\theta}_j)$ do not need to stem from the same distribution family. The cumulative density function (CDF) $G_j(\cdot, \boldsymbol{\theta}_j)$ of $P_j(\cdot, \boldsymbol{\theta}_j)$ allows to transform the measurements for feature $\tilde{\boldsymbol{Z}}_{,j}$ such that they follow a uniform distribution on the interval $(0, 1)$: $G'_j \equiv G'_j(\tilde{\boldsymbol{Z}}_{,j} - 1, \theta_j) \sim U(0, 1)$. Gaussian copula [153], the variant used in the scDesign family, then use the inverse CDF $F^{-1}(\cdot)$ of a standard normal distribution to obtain a transformed data matrix $\boldsymbol{W}$ where every feature follows a marginal Gaussian

distribution:

$$\boldsymbol{W}_{,j} = F^{-1}(G'_j). \tag{6.2}$$

If $\tilde{\boldsymbol{Z}}$ is discrete, for example in the case of gene expression data, G'_j follows a discrete uniform distribution. Therefore, G'_j must be distorted to be continuously uniform in order to obtain a continuous-valued $\boldsymbol{W}$ in Equation 6.2. In scDesign2 [156], this is achieved through the distributional transform $G^*_j = \boldsymbol{v}_j G'_j + (1 - \boldsymbol{v}_j)G'_j$, where each $\boldsymbol{v}_j$ is randomly sampled from a m-dimensional uniform distribution [158].
Simulating new samples with the same correlation structure as the transformed data in the Gaussian space is now possible through sampling from a multivariate normal distribution: $\hat{\boldsymbol{W}} \sim N_r(0, \boldsymbol{R})$. Here, $\boldsymbol{R}$ is the empirical correlation matrix of $(F^{-1}(G'_1), \ldots, F^{-1}(G'_r))^T$ or $(F^{-1}(G^*_1), \ldots, F^{-1}(G^*_r))^T$, respectively. Finally, reversing the transformation from before, i.e applying a standard Gaussian CDF followed by the inverse CDF of the respective marginal feature $j = 1, \ldots, r$ yields $\hat{\boldsymbol{Z}}_{,j} = G_j^{-1}(F(\hat{\boldsymbol{W}}_{,j}), \theta_j)$. Through this process synthetic data $\hat{\boldsymbol{Z}}$ with the same marginal feature distributions and correlation structure as $\tilde{\boldsymbol{Z}}$ can be generated.
In the `BacSC` pipeline (Contribution [4], Appendix C.1), I used the scDEED method for determining hyperparameters in UMAP embeddings [148], which uses a randomized negative control dataset. For false discovery rate control in differential gene expression testing, I employed the ClusterDE method [152], which relies on a synthetically generated null dataset through scDesign [156, 157].

6.3. Data Decomposition and Synthetic Null Generation in the BacSC Pipeline

With the recent emergence of different protocols for bacterial scRNA-seq [29, 30, 31, 32, 33, 34], a gold standard for statistical processing of the gene expression data produced by these methods is essential. Due to the differences between eukaryotic and bacterial scRNA-seq protocols described in Section 1.1, gene expression matrices obtained from these procedures can heavily differ in their statistical properties such as dimensionality, sparsity, or overdispersion [4]. Because of this, statistical processing of bacterial scRNA-seq data must be adapted to the characteristics of the data at hand. In contribution [4] (Appendix C.1), I developed `BacSC`, a computational pipeline for quality control, variance stabilization, dimension reduction, embedding, clustering, and differential expression testing (see Figure 1.2C) of bacterial scRNA-seq data. Its main objective is to allow researchers to perform statistically valid processing of bacterial scRNA-seq data without requiring extensive knowledge about details of the sequencing protocol and theory behind the statistical methods. To this end, important hyperparameters such as the latent rank of a singular value decomposition, the number of neighbors in UMAP embeddings [26], or the resolution of a cell type clustering [141, 142] are chosen automatically based on the characteristics of the data. Throughout the `BacSC` pipeline, I repeatedly used count splitting and synthetic null generation for automatic, data-driven determination of hyperparameters and FDR correction in hypothesis testing. This section gives a short overview, explaining how I applied the techniques described above in each task.

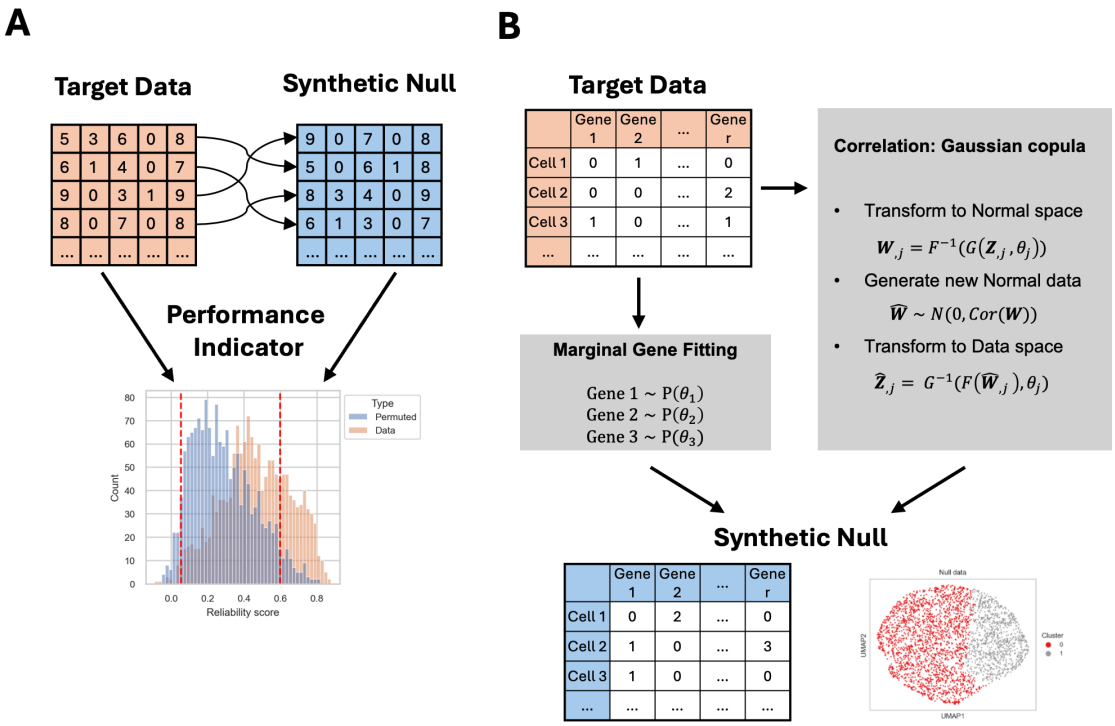


Figure 6.2.: **Synthetic null data generation approaches. A)** Randomization can be used to obtain a null distribution for a performance indicator. Comparing the quantiles of the null and data performance indicator distributions (the dashed red lines denote the 5% and 95% quantiles of the null distribution) indicates whether the method is suited to the dataset at hand. **B)** Schematic representation of a Gaussian copula approach to generate synthetic null data with marginal distributions correlation structure equal to the original data. *Figure adapted from Figures 1 and B1 in contribution [4].*

Before processing the data, `BacSC` first detects and removes cells that do not contain a sufficient number of reads, as well as outliers that are likely caused by measurement errors, resulting in a raw gene expression matrix $\tilde{\boldsymbol{Z}} \in \mathbb{N}_0^{m \times r}$ with m cells and r genes [5]. Next, the read counts are variance-stabilized through the log-transformation with feature-specific pseudocounts, as described in Section 5.2 [129], and subsequently scaled to have zero mean and unit variance per gene, which yields a variance-stabilized gene expression matrix $\boldsymbol{Z} \in \mathbb{R}^{m \times r}$.

Dimensionality Reduction. The dimensionality reduction step uses the count splitting approach proposed by Neufeld et al. [140] to select the number of latent dimensions in a singular value decomposition (SVD) of $\boldsymbol{Z}$. To this end, `BacSC` first determines whether the raw gene expression data $\tilde{\boldsymbol{Z}}$ follows a Poisson- or Negative-Binomial distribution. The respective count splitting algorithm then constructs training and test datasets, $\tilde{\boldsymbol{Z}}_{train}$ and $\tilde{\boldsymbol{Z}}_{test}$, that both contain information on all cells and genes (Figure 6.3A, panel 1 and 2). Both training and test data are variance-stabilized and scaled as described above, yielding

$\boldsymbol{Z}_{train}$ and $\boldsymbol{Z}_{test}$. Based on the SVD of the training data, $\boldsymbol{Z}_{train} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, the latent dimensionality k_{opt} is determined by minimizing the difference between k-dimensional truncations of the training data's SVD and the test data (Figure 6.3A, panel 3) over a range $k = 1, \ldots, k^*$ of possible dimensionalities:

$$L_k = ||\boldsymbol{Z}_{test} - \boldsymbol{U}_{\cdot,1:k}\boldsymbol{\Sigma}_{1:k,1:k}\boldsymbol{V}^T_{\cdot,1:k}||^2_F \tag{6.3}$$

$$k_{opt} = \underset{k=1,\ldots,k^*}{\arg\min}\, L_k. \tag{6.4}$$

UMAP Embedding. UMAP embeddings [26] visualize the gene expression data in two dimensions as a point cloud (Figure 6.3B, panel 3). To this end, a neighborhood graph detailing the pairwise distances between each cell and its closest neighbors in the k_{opt}-dimensional truncated SVD of $\boldsymbol{Z}$ is constructed. The two most important hyperparameters for UMAPs are $n_{neighbors}$, indicating the number of neighbors considered for each cell, and min_{dist}, the minimal distance between cells in the visualization. These should be chosen such that that the distances between similar cells are preserved, while cells with different gene expression profiles are not placed close to each other. The scDEED method [148], which I employ in the `BacSC` pipeline, uses a correlation-based approach to evaluate this property, categorizing cell embeddings as trustworthy or dubious (Figure 6.3B, panel 3) based on how much the UMAP embedding perturbs distances in the neighborhood of the cell. Since these correlation patterns are highly dependent on the dataset, no generally applicable null distribution can be defined. Instead, the correlation values are compared to the quantiles of a randomized synthetic null dataset (Figure 6.3B, panel 1), calculated by randomly shuffling the entries in the distance matrix as described in Section 6.2.

Clustering. In `BacSC`, I use count splitting to select a suitable resolution parameter in a Leiden [142] clustering. For a given resolution parameter γ, I first determine the cluster assignment on the training data through Leiden clustering, which maximizes the modularity:

$$M(\gamma) = \frac{1}{2e_t}\sum_c (e_c - \gamma\frac{K_c^2}{2m}). \tag{6.5}$$

The modularity is defined in terms of the neighborhood graph used in the UMAP embedding, where e_t is the total number of edges in the neighborhood graph, e_c is the number of edges within cluster c, and K_c is the sum of degrees over all nodes in cluster c. Because count splitting preserves the cell identities, it is valid to calculate the modularity $M_{test}(\gamma)$ of the cluster assignment obtained from the training data on the test data (Figure 6.3C, Panels 1, 2) as a measure of generalization power for γ. As the modularity depends on the number of clusters in the data, it will be maximal if γ is minimal, i.e. all data points are assigned to the same cluster, making simple test modularity maximization infeasible. By calculating the modularity of another random cluster assignment on the test data with the same number of clusters and size proportions as the the solution found for γ, I obtain a clustering-specific baseline. The best resolution γ_{opt} is then chosen to maximize the improvement of $M_{test}(\gamma)$ over the modularity of the corresponding random cluster assignment $M_{random}(\gamma)$ (Figure 6.3C, Panel 3):

$$\gamma_{opt} = \underset{\gamma=0.01,\ldots,0.5}{\arg\max} \left(M_{test}(\gamma) - M_{random}(\gamma)\right). \tag{6.6}$$

The method is easily adaptable to hyperparameters in other clustering algorithms, such as the perplexity Louvain clustering [141], by adapting the range of possible parameter values and the evaluation metric.

Differential Expression Testing. Differential expression testing determines characteristic genes that distinguish each of the cell type clusters from the rest of the cell population (Figure 6.3D, Panel 1) [17]. Testing for differences in gene expression between cell clusters that were defined through the very same gene expression profiles can lead to an inflated false discovery rate though, a phenomenon that is also known as "double dipping" [159, 160]. To avoid inflated false discovery rates, I use an adapted version of the ClusterDE method [152] in `BacSC`. ClusterDE combats double dipping through a strategy similar to permutation testing, contrasting the p-values obtained by testing each gene in $\boldsymbol{Z}$ for differential expression with p-values obtained on a synthetic null. To this end, a synthetic null dataset that has the same marginal moments as the gene expression data but no apparent cluster structure (Figure 6.3D, Panel 2) is simulated through a copula approach adapted from scDesign2 [156]. Forcing a clustering on the synthetic null data can now provide a distribution of p-values that is associated with false discoveries and thus serve as a baseline to control the FDR (Figure 6.3D, Panel 2). Finally, the clipper method [161] calculates contrast scores between the real and synthetic p-values, and yields a threshold that provides FDR control (Figure 6.3D, Panel 3).

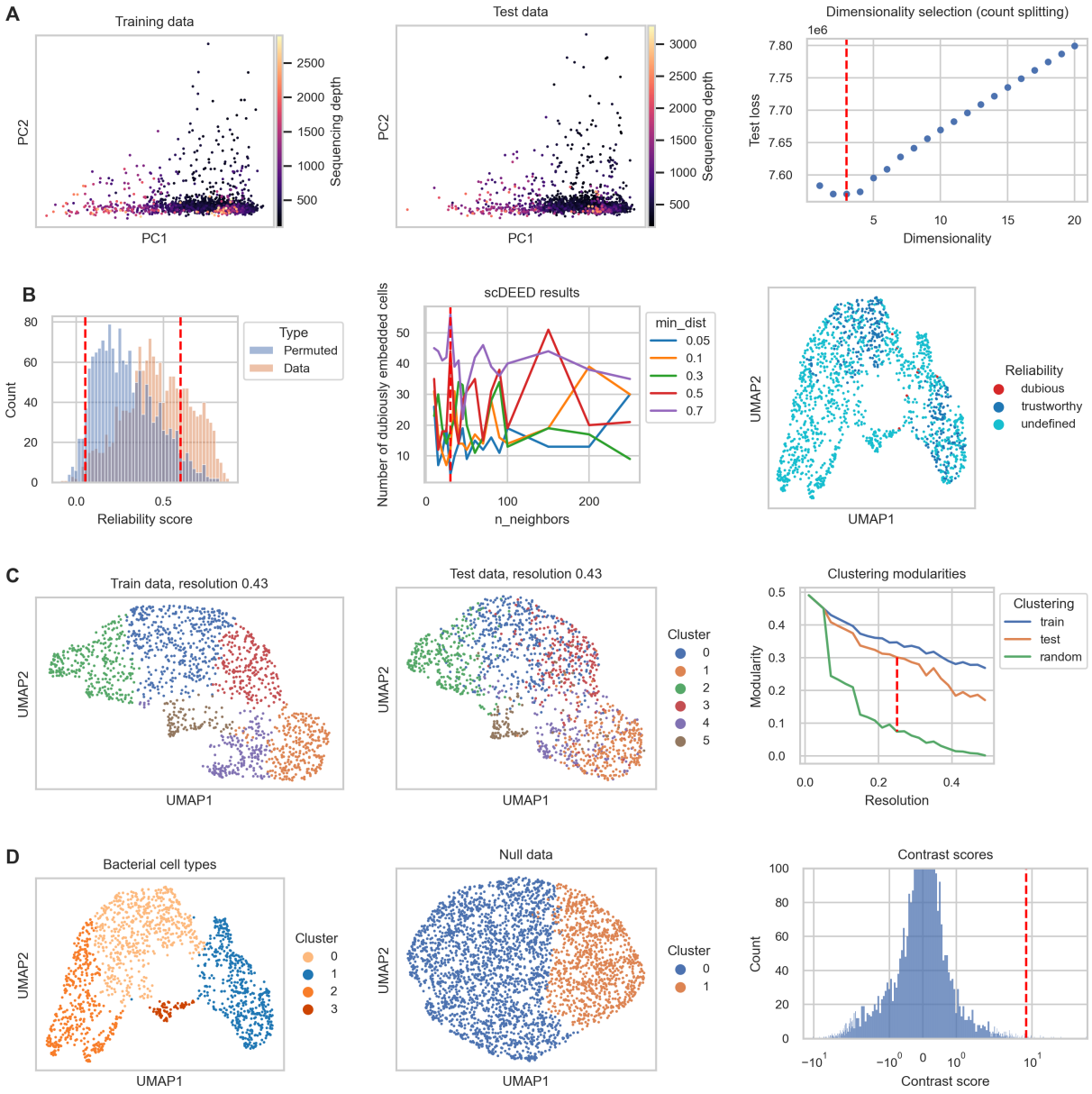


Figure 6.3.: **Core concepts in the BacSC pipeline. A)** Count splitting divides the data into train and test data and allows to determine the latent dimensionality through count splitting. The dashed line in the third panel indicates the chosen number of dimensions. **B)** Panel 1: Distribution of reliability scores and classification cutoffs in scDEED. Panel 2: Selecting the parameters with the least number of dubiously embedded cells yields the UMAP in panel 3. **C)** Panel 1/2: Applying the training data clustering to the test data. Panel 3: Modularities of training, test, and random clustering with chosen resolution as a dashed line. **D)** Panel 1/2: UMAP embeddings of real and synthetic null data with colored clusters. Panel 3: Distribution of contrast scores with differential expression cutoff. *All panels are part of Figures B1-B3 in contribution [4].*

7. Summary of the contributions

In this chapter, I give an overview over the five contributions (Appendices A - C) that are part of this thesis and outline my contribution to each project. To this end, I connect the ideas presented in chapters 2-6 to the specific methods developed in the contributions and highlight concrete applications on experimental data. Overall, the contributions in this thesis provide new methods and results for the statistical analysis of sparse count data generated by high-throughput sequencing protocols. The contributions focus on the analysis of different data modalities from single-cell RNA and amplicon sequencing, in particular differential analysis of compositional cell type abundances and processing of gene expression data.

Generative modeling and testing for differences in compositional abundance data, therefore deriving statistically accurate and biologically meaningful descriptions of perturbations in cellular compositions, constitutes the first area of contributions, which includes the `scCODA` ([1], see Appendix A.1), `tascCODA` ([2], see Appendix A.2), and `cosmoDA` ([3], see Appendix B) models. I mostly restricted my focus to datasets with a low or medium dimensionality, opening up the possibility to pay more attention to effective parameter estimation and flexible experimental designs. The compositional modeling toolbox described in chapters 2 - 5 provides an overview over the statistical approaches used in `scCODA`, `tascCODA`, and `cosmoDA`. Each contribution uses a different set of techniques from this toolbox, and additionally focuses on incorporating other sources of information into the modeling and differential testing process (see Table 7.1 for an overview).

	`scCODA` [1]	`tascCODA` [2]	`cosmoDA` [3]
Distribution	Dirichlet	Dirichlet	a-b power interaction model
Parameter estimation method	Hamiltonian Monte Carlo	Hamiltonian Monte Carlo (NUTS)	Score Matching
Model selection/DA Testing	Spike-and-slab priors	Spike-and-slab LASSO priors	Hypothesis test
Count data	Multinomial	Multinomial	Relative abundance
Zero handling	Constant imputation	Constant imputation	Power transformation
Other features	Specialized for scRNA-seq data	Tree-adaptive DA testing	Feature associations

Table 7.1.: Overview over concepts used in differential abundance testing methods developed in contributions [1], [2], and [3].

Contribution 1 ([1], Appendix A.1). This contribution presents the `scCODA` model, a generative, hierarchical Bayesian model for compositional differential analysis of cell type abundance data from scRNA-seq. At its base, `scCODA` uses a Dirichlet-Multinomial distribution to model the compositional cell type counts (y_i in Figure 7.1A, Table 7.1). The abundance of each cell type is connected to the covariates of interest X_i through a log-linear model with intercepts α_k and effects β_k (see Section 4.1). Through the use of continuous spike-and-slab priors as described in Section 4.2.2 and thresholding of the posterior inclusion probability (Section 4.3), statistically credible effects on β_k are selected. In its initial implementation, `scCODA` used Hamiltonian Monte Carlo sampling (Section 3.1) for parameter estimation. I further enhanced the computational efficiency through No-U-Turn sampling in a later reimplementation [162]. To further improve the convergence rate of the HMC sampler, zero entries can be replaced by a constant value (see Section 5.2), while automatic selection of a reference cell type that is assumed to have no credible effects guarantees identifiability of the result.
This article is the first work to acknowledge the compositionality of scRNA-seq cell type abundance data, showing the inadequacy of non-compositional tests for differential abundance on simulated and real scRNA-seq cell type abundances. In extensive synthetic data simulations, I showed that `scCODA` achieved superior performance especially in low sample-size situations, which are characteristic for scRNA-seq experiments (Figure 7.1B). Notably, `scCODA` was the only method to adequately control the false discovery rate during simulations, even outperforming established methods for compositional differential abundance testing of amplicon sequencing data like ANCOM-BC [107] and ALDEx2 [98]. Applications of `scCODA` presented in [1] include the detection of decreased B-cell abundances in peripheral blood mononuclear cells (PBMCs) of supercentenarians [77], description of diverse changes in the cell type composition of the intestinal epithelium and lamina propria of subjects with ulcerative colitis [72]. An especially instructive example is shown in Figure 7.1C, analyzing the intestinal epithelium of mice under infections with *Salmonella* and *Heligmosomoides polygyrus* [73]. Here, `scCODA` detected significantly less cell types as differentially abundant than the originally used Poisson regression model, suggesting that these additional discoveries are false-positive results.
I started work on `scCODA` during my Master's thesis under the supervision of Dr. Benjamin Schubert, Dr. Maren Büttner, and Prof. Dr. Fabian Theis, developing a preliminary version of the model. During my doctoral studies, I refined the statistical methodology by adding the algorithms for automatic reference selection, adequate FDR control, and credible interval calculation with input from Dr. Benjamin Schubert, Dr. Maren Büttner, and Prof. Dr. Christian L. Müller. During this period, I also conducted the benchmarks for model comparison, heterogeneous response groups, and runtime analysis, which were conceived together with the same co-authors, as well as the data application on bacterial infections of the intestinal epithelium of mice. I also finalized the implementation of a Python package for `scCODA` during my PhD studies, and maintained the package over the following years. I was further responsible for the sections on model description and benchmarking performance in the main text and supplement of the article.

Contribution 2 ([2], Appendix A.2). In this contribution, I developed `tascCODA`, an extension of `scCODA` for tree-aggregated differential abundance testing of general high-throughput sequencing abundance data. As described in Section 1.2, scRNA-seq cell types

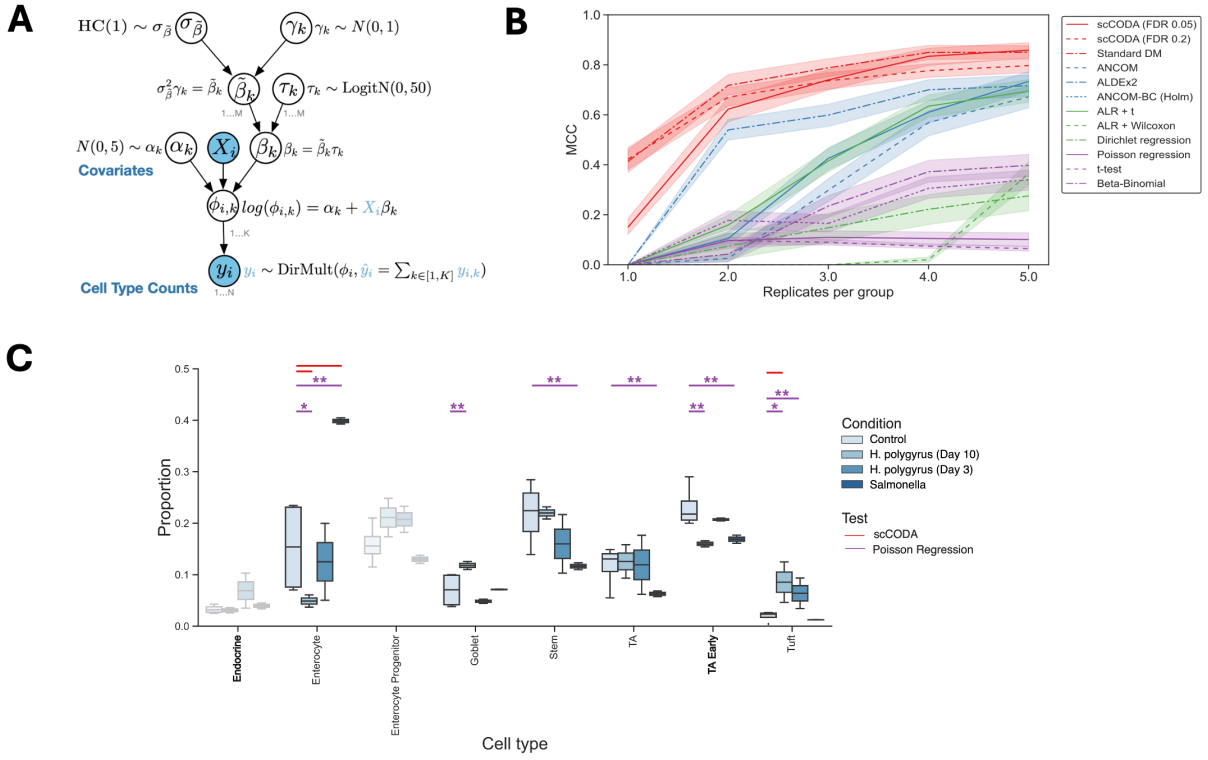


Figure 7.1.: **The scCODA model for compositional analysis of scRNA-seq data [1]. A)** Schematic representation of the hierarchical Bayesian formulation. **B)** Matthews correlation coefficient (MCC) of scCODA and other methods on simulated data. **C)** Cell type compositions of the intestinal epithelium of mice under different bacterial infections [73]. The colored horizontal lines denote differentially abundant cell types detected by scCODA and Poisson regression. *Figure adapted from contribution [1].*

and OTUs/ASVs/taxa in amplicon sequencing share many statistical properties and are usually ordered hierarchically through cell lineage, phylogenetic, or taxonomic trees. Like `scCODA`, `tascCODA` uses a Dirichlet-Multinomial model for describing compositional count data (y in Figure 7.2A, Table 7.1). By adding additional effects on the internal nodes of the tree (e.g. N3 in Figure 7.2B), `tascCODA` can aggregate entire groups of features if they are effected by a covariate in the same way. An adaptive regularization scheme using spike-and-slab LASSO priors (Sections 4.2.2, 5.3) with regularization strengths λ_0 and λ_1, respectively, ensures model identifiability. Through a user-defined hyperparameter ϕ, the penalization can be adjusted to prefer general effects near the root of the tree or more detailed effects near the leaf nodes. This model selection scheme, paired with a suitable beta prior on the mixture component θ and thresholding of the posterior inclusion probability to control the FDR (see Section 4.3), also allows for differential abundance testing. As with `scCODA`, parameter estimation for `tascCODA` was performed through HMC sampling in the initial implementation, replaced by NUTS sampling (Section 3.1) in a later reimplementation [162]. For handling of zero entries and reference feature

selection, `tascCODA` also uses the same strategies as `scCODA`.
In the synthetic data simulations presented in contribution [2], `tascCODA` was able to outperform other established DA testing methods if a detailed aggregation level was chosen, and achieved adequate results if higher aggregations were preferred (Figure 7.2C). One application on real data analyzes scRNA-seq data from the intestinal epithelium and lamina propria of subjects with ulcerative colitis and healthy controls [72], where I showed that biasing the model selection in `tascCODA` towards the leaves of the tree gradually leads to a less sparse solution with better out-of-sample prediction performance. For an application on microbial abundance data, I analyzed changes in the gut microbial composition among patients with irritable bowel syndrome [163] (Figure 7.2D). Here, `tascCODA` gave similar results as `scCODA`, but was able to additionally detect aggregated effects as well.
For this project, I conceived the statistical model and designed all simulations and out-of-sample prediction studies with suggestions from Prof. Dr. Christian L. Müller. I was responsible for the implementation of tascCODA, conducted all simulation studies and data applications, except for the initial processing of microbiome data, which was performed by Salomé Carcy. The manuscript was written by me with help from Prof. Dr. Christian L. Müller. I further developed and maintained the accompanying software package for Python.

Contribution 3 ([3], Appendix B). The influence of pairwise feature associations and zero replacement are two other rarely discussed topics in the context of compositional DA testing. This gap is filled by the `cosmoDA` model developed in this contribution. If features interact with each other, i.e. the red and yellow features in Figure 7.3A, their abundances will be strongly correlated beyond the compositional effect. Differential abundance testing must take these associations into account to avoid false positive results caused by secondary effects. These are not a direct result of the covariate, but occur on features that have a strong correlation with a differentially abundant feature (see Figure 7.3A). The `cosmoDA` model uses a-b power interaction models (Section 2.1.5) to estimate feature associations and thus account for their impact when determining covariate-induced effects $\boldsymbol{\eta}_1$ through a linear model on the location vector (Figure 7.3A, Table 7.1). Through an extension of the score matching estimator for a-b power interaction models ([42], Section 3.2), efficient parameter estimation of this covariate-extended model is possible. To avoid model misspecification, `cosmoDA` uses LASSO regularization on the off-diagonal entries of the interaction matrix (see Section 4.2.1) and selects the regularization strength through cross validation. Differential abundance on $\boldsymbol{\eta}_1$ is determined through hypothesis testing, as described in Section 4.3. Furthermore, counts are simply transformed to relative abundances by taking the closure (Equation 1.2). As discussed in Section 5.2, the data transformation used in a-b power interaction models is closely related to Box-Cox transformations and therefore able to avoid the need for zero replacement. Because of this, `cosmoDA` does not rely on a zero imputation strategy, and can instead work with unperturbed relative abundance data.
As with the previous models, I evaluated `cosmoDA` on synthetic and real high-throughput sequencing abundance data. In a simulated data benchmark with correlated features, `cosmoDA` showed a better overall performance than other methods, including ANCOM-BC [107] and CompDA [112], a method that also takes feature associations into account (Figure 7.3B). Especially when the sample size n was larger, `cosmoDA` also showed good FDR

control, although the desired level was not always met. On experimental scRNA-seq data comparing PBMCs from healthy controls and subjects with systemic lupus erythematosus (SLE) [74], `cosmoDA` was the only tested method that did not show a significant association of SLE with clonal monocyte abundance, a result that was confirmed by blood count samples [74]. I also investigated the impact of the power transformation on the differential abundance results. Comparing microbial abundances in the gut microbiomes obtained by 16S rRNA sequencing from infants in Malawi and the United States [67] showed varying differential abundance results for different exponents ϕ in the power transformation (Figure 7.3C). The use of a pseudocount for zero replacement also had considerable impact on the set of differentially abundant phyla. Notably, results for ANCOM-BC [107], which implicitly adds pseudocounts to zero entries, and `cosmoDA` with added pseudocounts and ϕ selected by Procrustes analysis (see Section 5.2) produced almost identical sets of DA phyla.
For this project, I conceived the `cosmoDA` model with input from Prof. Dr. Hongzhe Li and Prof. Dr. Christian L. Müller. I developed the model, reimplemented and extended the genscore package for R [64] in Python, and was responsible for design, execution, and evaluation of all simulation studies and data applications. I further wrote the manuscript with suggestions from Prof. Dr. Hongzhe Li and Prof. Dr. Christian L. Müller.

The second objective of this thesis, fulfilled by contributions [4] and [5], encompasses the establishment of best practices and statistically sound pipelines for scRNA-seq data analysis.

Contribution 4 ([4], Appendix C.1). With the `BacSC` pipeline described in this contribution, I facilitate processing of bacterial scRNA-seq gene expression data with only minimal manual intervention or expert knowledge required, automatically adapting to the characteristics of datasets from various sequencing protocols. The pipeline starts with a quality control step to remove empty droplets, doublets, and other outlier measurements, before performing a variance-stabilizing transform and scaling step (see Sections 5.2, 6.3). Using the techniques and methods introduced in Chapter 6, `BacSC` provides solutions for automatic selection of important hyperparameters in three steps of the scRNA-seq data processing pipeline:

- The latent dimensionality of a singular-value decomposition of a variance-stabilized gene expression dataset through count splitting ([140], Section 6.1).
- The number of neighbors in the neighborhood graph and minimal distance in a UMAP embedding [26] through scDEED [148].
- The resolution of a Leiden [142] or Louvain [141] clustering through a count-splitting approach.

Finally, `BacSC` facilitates valid differential expression testing and subsequent cell type annotation by correcting for inflated false discovery rates caused by double dipping. For this, I adapted the ClusterDE method [152] to highly sparse gene expression values and skewed cluster proportions. I discussed the methodology used in the individual steps in more detail in Section 6.3.

To show the broad application range of `BacSC`, contribution [4] contains applications on 13 different datasets, stemming from two different bacterial scRNA-seq protocols [29, 30] and containing cells from five different species. For *Bacillus subtilis* grown in minimal media and sequenced with the ProBac-seq protocol [30, 31], `BacSC` discovered five different cell types, of which some were clearly identifiable as competent cells, spores, and cells with increased expression of genes relating to structural flagella components after differential expression testing (Figure 7.4A). Notably, the UMAP embedding shows continuous transitional states between the cell types, which were not picked up by previous analyses of the same dataset [30]. Analyzing populations of *Klebsiella Pneumoniae* subjected to three different antibiotics and sequenced with the BacDrop protocol [29], `BacSC` found clear, replicable separation between the treatments and discovered a population of mobile genetic elements that was also reported previously (Figure 7.4B). I conducted another showcase of data integration from different growth conditions on *Pseudomonas aeruginosa* grown in regular and low-iron environments and sequenced with ProBac-seq. Here, `BacSC` found cell type clusters spanning both environments, as well as between-condition differences in gene expression (Figure 7.4C, panel 1). The set of differentially expressed (DE) genes between the two conditions found on the scRNA-seq data had considerable overlap with DE genes detected by different methods on a comparable bulk sequencing experiment from the Co-PATHOgenex study [164] 7.4 C, panel 2). Furthermore, most of the genes differentially expressed in both datasets (e.g. PA4514, icmP, phuR, Figure 7.4C, panel 3) are known to be related to iron reception.
This project was carried out in collaboration with Tim Kirk, Dr. Janne Gesine Thöming, Prof. Dr. Susanne Häußler, and Prof. Dr. Adam Z. Rosenthal, who provided all datasets generated with the ProBac-Seq protocol. With suggestions from Prof. Dr. Christian L. Müller, I designed the structure and individual steps of the BacSC pipeline. I conducted its applications to all datasets and evaluated the results with help from Tim Kirk, Dr. Janne Gesine Thöming, Prof. Dr. Adam Z. Rosenthal, and Roberto Olayo Alarcon. I was further responsible for the manuscript, designed all figures and wrote all text with the exception of the section on biological data generation, which was written by Tim Kirk, Dr. Janne Gesine Thöming, and Prof. Dr. Adam Z. Rosenthal.

Contribution 5 ([5], Appendix C.2). This contribution provides guidelines and computational pipelines for the statistically sound analysis of single-cell RNA sequencing data. It serves as an updated and extended version of the best-practice recommendations given by Lücken and Theis [17] and gives a broad overview over tools and methods for every step of the scRNA-seq analysis pipeline (see Figure 1.2C), as well as advanced techniques such as spatial analysis, diffusion pseudotime, or multi-omics analysis. To simplify the application of these recommendations and allow for constant updating with new methods and topics, the paper is tied to an online book that consists of interactive notebooks, detailing every analysis step with code examples on real data.
As a member of the Single Cell best practices consortium, I contributed to the section on compositional data analysis of cell-type abundance data, covering differential abundance testing with `scCODA` and `tascCODA` (Contributions [1] and [2]) and their aforementioned reimplementations in the `pertpy` package for Python [162]. I wrote the corresponding notebook and section in the publication with suggestions from Lukas Heumos and Prof. Dr. Christian L. Müller.

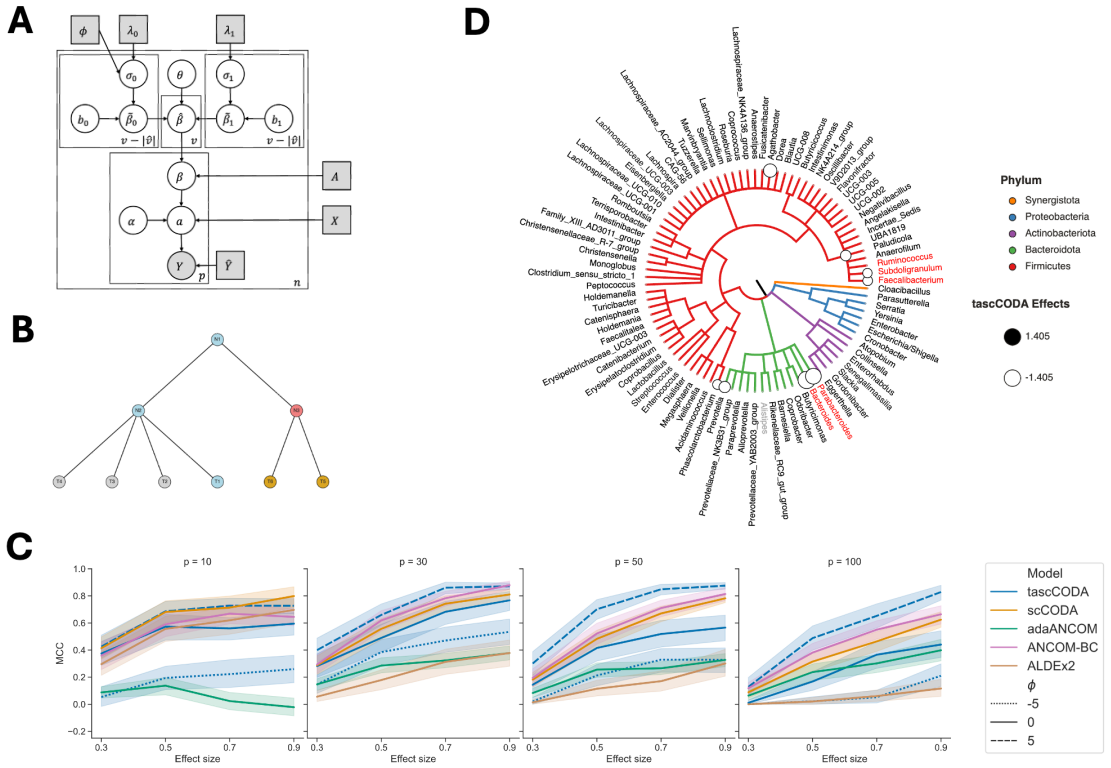


Figure 7.2.: **The tascCODA model for tree-aggregated compositional analysis of HTS data [2]. A)** Schematic representation of the hierarchical Bayesian formulation. **B)** Example tree structure with internal nodes N1, N2, N3, and leaves T1, . . . , T6. Joint effects on leaf nodes (e.g. N5 and N6) can be represented by an aggregated effect on the corresponding internal node (N3). **C)** Matthews correlation coefficient (MCC) of tascCODA and other methods on simulated data. **D)** Credible changes found by tascCODA, comparing healthy controls and IBS patients in the genus-aggregated data of [163]. The circles on nodes of the tree represent credible effects, red genera show the credible effects found by scCODA (FDR 0.1) on the genus level. The gray genus *Alistipes* was used as the reference for tascCODA and scCODA. *Figure adapted from contribution [2].*

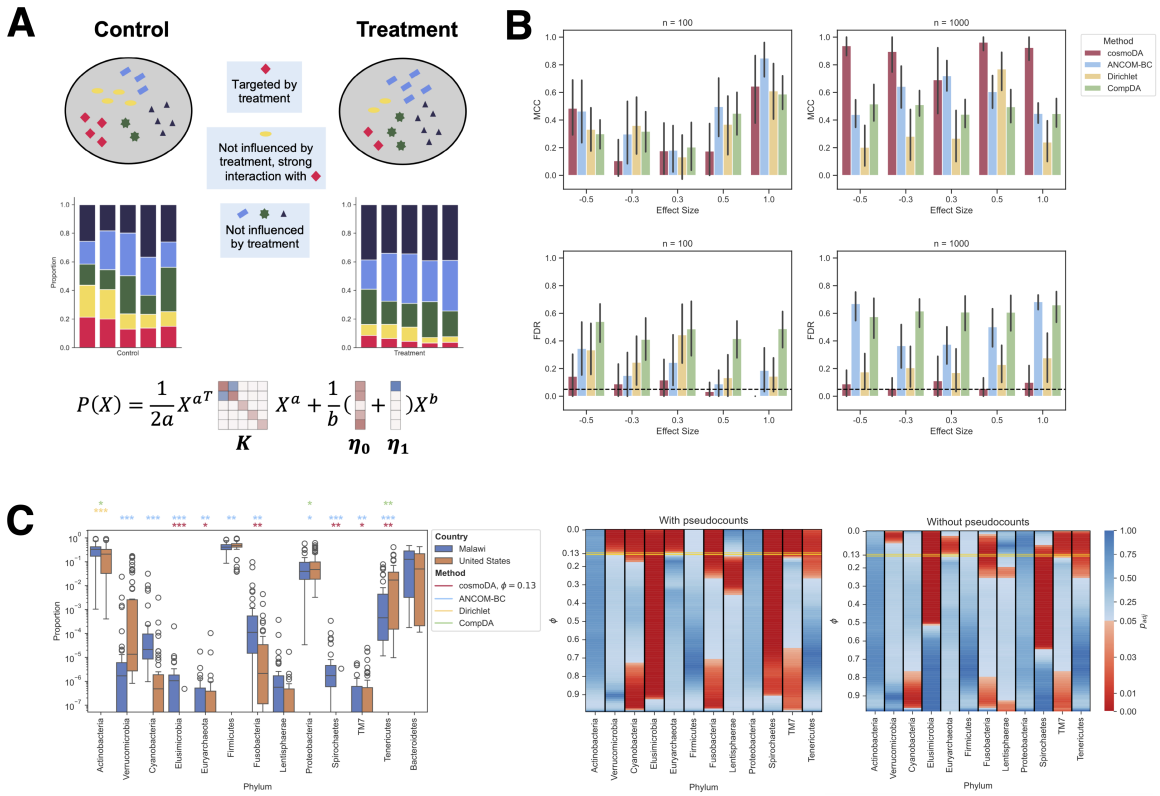


Figure 7.3.: **The cosmoDA model for compositional differential abundance analysis of HTS data with interactions [3]. A)** Distinction between primary (red) and secondary (yellow) effects. Through modeling feature interactions, cosmoDA can distinguish between the two effect types. **B)** Matthews correlation coefficient (MCC) and false discovery rate (FDR) of cosmoDA and other methods on simulated data. **C)** Differential analysis of infant gut microbiota in Malawi and the US. Left panel: Boxplots of relative phylum abundances. The stars indicate adjusted p-values for different methods (*: $p_{adj} < 0.05$; **: $p_{adj} < 0.01$; ***: $p_{adj} < 0.001$). Middle and left panel: Adjusted p-values for cosmoDA with different power transformations. The yellow box highlights the adjusted p-values for ϕ determined through Procrustes correlation analysis. *Figure adapted from contribution [3].*

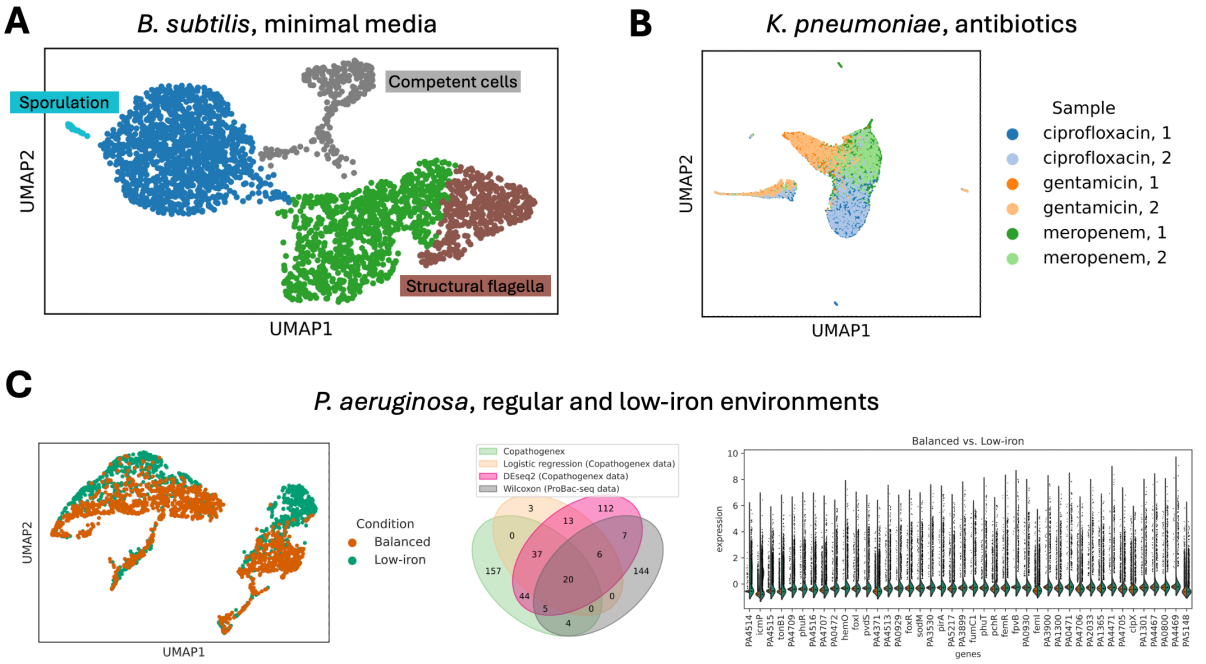


Figure 7.4.: **Key findings from processing bacterial scRNA-seq data with the BacSC pipeline [4]. A)** UMAP plot of *B. subtilis* grown in minimal media after processing with BacSC. **B)** UMAP plot of *K. pneumoniae* after treatment with different antibiotics and scRNA-seq data processing with BacSC. **C)** Joint analysis of *P. aeruginosa* grown in low-iron and regular environments with BacSC. Left panel: UMAP plot after processing with BacSC. Middle panel: Venn diagram of differentially expressed genes found in Co-PATHOgenex and ProBac-seq data. Right panel: Violin plots of differentially expressed genes in ProBac-seq and Co-PATHOgenex data. *Figure adapted from contribution [4].*

8. Outlook

Thinking beyond the ideas presented throughout this thesis, there are multiple possible directions for future research.
One major area of interest is the full unification of the compositional distributions in Chapter 2. While this thesis describes an overarching distribution family using a-b power interaction models (PIM), understanding the constraints required to achieve full equivalence between individual distributions similar to [39] is a crucial next step to fully bridge the gaps between these approaches for compositional modeling. Formally connecting the parameters of the PIM to the moments of the compositional distribution can further lead to more exact interpretations than the ones given in contribution [3].
Another possible area of extension are more complex experimental setups, such as settings with multiple covariates, mixed-effect models, or longitudinal data [165, 166]. Simultaneous differential abundance testing on multiple covariates is already possible for the `scCODA` and `tascCODA` models, while the `cosmoDA` model, as presented in contribution [3], only supports a single covariate. Extending the linear model formulation from (Equation 4.1), which is used by all three methods, is however straightforward. Possible generalizations can include generalized linear models [167] of the form $\boldsymbol{\eta} = g^{-1}(\boldsymbol{Y}\boldsymbol{\psi})$, or even generalized additive models [168] of the form

$$\boldsymbol{\eta} = \boldsymbol{\eta}_0 + \sum_{l=1}^{d} f_l(\boldsymbol{Y}_{,l}). \tag{8.1}$$

Such extensions are however tied to the development of more flexible parameter estimation approaches. In score matching optimization, automatic differentiation algorithms can provide the desired flexibility [169], while hierarchical Bayesian approaches would require additional prior structures to model more complex experimental settings. In the face of constantly increasing sample sizes for all kinds of HTS experiments, even interpretable deep learning approaches [170, 171], augmenting differential abundance testing with nonlinear estimation of more complex sources of variation, are imaginable.
On the side of best practices for bacterial scRNA-seq data analysis, a more comprehensive evaluation of other methods than the (relatively basic) ones used in the individual steps of `BacSC` could yield even better results than the relatively standard tools used so far. Considering the enormous and ever-increasing number and variety of available methods [5, 172], a full evaluation and comparison of all available tools may however be a complicated and time-consuming task.
Finally, multi-omics datasets, jointly measuring multiple sources of information such as genomics, transcriptomics, metabolomics, proteomics, or epigenomics, allow for even more detailed analysis of cellular populations [173]. Adapting the methods and frameworks presented here to consider these additional sources of information is therefore crucial to ensure their longevity.

Bibliography

[1] Büttner, M., Ostner, J., Müller, C.L., Theis, F.J., Schubert, B.: scCODA is a bayesian model for compositional single-cell data analysis. Nat. Commun. **12**(1), 6876 (2021) https://doi.org/10.1038/s41467-021-27150-6

[2] Ostner, J., Carcy, S., Müller, C.L.: tascCODA: Bayesian tree-aggregated analysis of compositional amplicon and single-cell data. Front. Genet. **12**, 766405 (2021) https://doi.org/10.3389/fgene.2021.766405

[3] Ostner, J., Li, H., Müller, C.L.: Score matching for differential abundance testing of compositional high-throughput sequencing data. bioRxiv (2024) https://doi.org/10.1101/2024.12.05.627006

[4] Ostner, J., Kirk, T., Olayo-Alarcon, R., Thöming, J., Rosenthal, A.Z., Häussler, S., Müller, C.L.: BacSC: A general workflow for bacterial single-cell RNA sequencing data analysis. bioRxiv (2024) https://doi.org/10.1101/2024.06.22.600071

[5] Heumos, L., Schaar, A.C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M.D., Strobl, D.C., Henao, J., Curion, F., Single-cell Best Practices Consortium, Schiller, H.B., Theis, F.J.: Best practices for single-cell analysis across modalities. Nat. Rev. Genet. **24**(8), 550–572 (2023) https://doi.org/10.1038/s41576-023-00586-w

[6] Haque, A., Engel, J., Teichmann, S.A., Lönnberg, T.: A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. **9**(1), 75 (2017) https://doi.org/10.1186/s13073-017-0467-4

[7] Xia, Y., Sun, J., Chen, D.-G.: Statistical Analysis of Microbiome Data with R. Springer, Singapore, Singapore (2018). https://doi.org/10.1007/978-981-13-1534-3 . https://link.springer.com/10.1007/978-981-13-1534-3

[8] Hugerth, L.W., Andersson, A.F.: Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing. Front. Microbiol. **8**, 1561 (2017) https://doi.org/10.3389/fmicb.2017.01561

[9] Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., B Tuch, B., Siddiqui, A., Lao, K., Azim Surani, M.: mRNA-seq whole-transcriptome analysis of a single cell. Nat. Methods **6**(5), 377–382 (2009) https://doi.org/10.1038/nmeth.1315

[10] Brennan, M.A., Rosenthal, A.Z.: Single-cell RNA sequencing elucidates the structure and organization of microbial communities. Front. Microbiol. **12**, 713128 (2021) https://doi.org/10.3389/fmicb.2021.713128

[11] Holmes, S., Huber, W.: Modern Statistics for Modern Biology. Cambridge University Press, Cambridge, England (2018)

[12] Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., Egozcue, J.J.: Microbiome datasets are compositional: And this is not optional. Front. Microbiol. **8**, 2224 (2017) https://doi.org/10.3389/fmicb.2017.02224

[13] Quinn, T.P., Erb, I., Richardson, M.F., Crowley, T.M.: Understanding sequencing data as compositions: an outlook and review. Bioinformatics **34**(16), 2870–2878 (2018) https://doi.org/10.1093/bioinformatics/bty175

[14] McGee, W.A., Pimentel, H., Pachter, L., Wu, J.Y.: Compositional data analysis is necessary for simulating and analyzing RNA-seq data. bioRxiv, 564955 (2019) https://doi.org/10.1101/564955

[15] Aitchison, J.: The statistical analysis of compositional data. J. R. Stat. Soc. Series B Stat. Methodol. **44**(2), 139–160 (1982)

[16] Mosimann, J.E.: On the compound multinomial distribution, the multivariate beta-distribution, and correlations among proportions. Biometrika **49**(1/2), 65–82 (1962)

[17] Luecken, M.D., Theis, F.J.: Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. **15**(6), 8746 (2019) https://doi.org/10.15252/msb.20188746

[18] Lin, H., Peddada, S.D.: Analysis of microbial compositions: a review of normalization and differential abundance analysis. NPJ Biofilms Microbiomes **6**(1), 60 (2020) https://doi.org/10.1038/s41522-020-00160-w

[19] Schuster, S.C.: Next-generation sequencing transforms today's biology. Nat. Methods **5**(1), 16–18 (2008) https://doi.org/10.1038/nmeth1156

[20] Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., Law, M.: Comparison of next-generation sequencing systems. J. Biomed. Biotechnol. **2012**(1), 251364 (2012) https://doi.org/10.1155/2012/251364

[21] Hu, T., Chitnis, N., Monos, D., Dinh, A.: Next-generation sequencing technologies: An overview. Hum. Immunol. **82**(11), 801–811 (2021) https://doi.org/10.1016/j.humimm.2021.02.012

[22] Kolbert, C.P., Persing, D.H.: Ribosomal DNA sequencing as a tool for identification of bacterial pathogens. Curr. Opin. Microbiol. **2**(3), 299–305 (1999) https://doi.org/10.1016/S1369-5274(99)80052-6

[23] Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., Teichmann, S.A.: The technology and biology of single-cell RNA sequencing. Mol. Cell **58**(4), 610–620 (2015) https://doi.org/10.1016/j.molcel.2015.04.005

[24] Hwang, B., Lee, J.H., Bang, D.: Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp. Mol. Med. **50**(8), 96 (2018) https://doi.org/10.1038/s12276-018-0071-8

[25] Andrews, T.S., Kiselev, V.Y., McCarthy, D., Hemberg, M.: Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. Nat. Protoc. **16**(1), 1–9 (2020) https://doi.org/10.1038/s41596-020-00409-w

[26] McInnes, L., Healy, J., Saul, N., Großberger, L.: UMAP: Uniform manifold approximation and projection. J. Open Source Softw. **3**(29), 861 (2018) https://doi.org/10.21105/joss.00861

[27] Walls, A.W., Rosenthal, A.Z.: Bacterial phenotypic heterogeneity through the lens of single-cell RNA sequencing. Transcription **15**(1-2), 48–62 (2024) https://doi.org/10.1080/21541264.2024.2334110

[28] Homberger, C., Barquist, L., Vogel, J.: Ushering in a new era of single-cell transcriptomics in bacteria. microLife **3** (2022) https://doi.org/10.1093/femsml/uqac020

[29] Ma, P., Amemiya, H.M., He, L.L., Gandhi, S.J., Nicol, R., Bhattacharyya, R.P., Smillie, C.S., Hung, D.T.: Bacterial droplet-based single-cell RNA-seq reveals antibiotic-associated heterogeneous cellular states. Cell (2023) https://doi.org/10.1016/j.cell.2023.01.002

[30] McNulty, R., Sritharan, D., Pahng, S.H., Meisch, J.P., Liu, S., Brennan, M.A., Saxer, G., Hormoz, S., Rosenthal, A.Z.: Probe-based bacterial single-cell RNA sequencing predicts toxin regulation. Nat Microbiol **8**(5), 934–945 (2023) https://doi.org/10.1038/s41564-023-01348-4

[31] Samanta, P., Cooke, S.F., McNulty, R., Hormoz, S., Rosenthal, A.: ProBac-seq, a bacterial single-cell RNA sequencing methodology using droplet microfluidics and large oligonucleotide probe sets. Nat. Protoc. (2024) https://doi.org/10.1038/s41596-024-01002-1

[32] Kuchina, A., Brettner, L.M., Paleologu, L., Roco, C.M., Rosenberg, A.B., Carignano, A., Kibler, R., Hirano, M., DePaolo, R.W., Seelig, G.: Microbial single-cell RNA sequencing by split-pool barcoding. Science **371**(6531) (2021) https://doi.org/10.1126/science.aba5257

[33] Imdahl, F., Vafadarnejad, E., Homberger, C., Saliba, A.-E., Vogel, J.: Single-cell RNA-sequencing reports growth-condition-specific global transcriptomes of individual bacteria. Nat Microbiol **5**(10), 1202–1206 (2020) https://doi.org/10.1038/s41564-020-0774-1

[34] Xu, Z., Wang, Y., Sheng, K., Rosenthal, R., Liu, N., Hua, X., Zhang, T., Chen, J., Song, M., Lv, Y., Zhang, S., Huang, Y., Wang, Z., Cao, T., Shen, Y., Jiang, Y., Yu, Y., Chen, Y., Guo, G., Yin, P., Weitz, D.A., Wang, Y.: Droplet-based

high-throughput single microbe RNA sequencing by smRandom-seq. Nat. Commun. **14**(1), 5130 (2023) https://doi.org/10.1038/s41467-023-40137-9

[35] Greenacre, M., Grunsky, E., Bacon-Shone, J., Erb, I., Quinn, T.: Aitchison's compositional data analysis 40 years on: A reappraisal. SSO Schweiz. Monatsschr. Zahnheilkd. **38**(3), 386–410 (2023) https://doi.org/10.1214/22-STS880

[36] Quinn, T.P., Erb, I., Gloor, G., Notredame, C., Richardson, M.F., Crowley, T.M.: A field guide for the compositional analysis of any-omics data. Gigascience **8**(9) (2019) https://doi.org/10.1093/gigascience/giz107

[37] Pearson, K.: Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc. R. Soc. Lond. **60**(359-367), 489–498 (1897) https://doi.org/10.1098/rspl.1896.0076

[38] Aitchison, J.: Principles of compositional data analysis. Lecture Notes-Monograph Series **24**, 73–81 (1994) https://doi.org/10.1214/LNMS/1215463786

[39] Aitchison, J.: A general class of distributions on the simplex. J. R. Stat. Soc. Series B Stat. Methodol. **47**(1), 136–146 (1985) https://doi.org/10.1111/j.2517-6161.1985.tb01341.x

[40] Scealy, J.L., Wood, A.T.A.: Score matching for compositional distributions. J. Am. Stat. Assoc. **118**(543), 1811–1823 (2022) https://doi.org/10.1080/01621459.2021.2016422

[41] Weistuch, C., Zhu, J., Deasy, J.O., Tannenbaum, A.R.: The maximum entropy principle for compositional data. BMC Bioinformatics **23**(1), 449 (2022) https://doi.org/10.1186/s12859-022-05007-z

[42] Yu, S., Drton, M., Shojaie, A.: Interaction models and generalized score matching for compositional data. In: Villar, S., Chamberlain, B. (eds.) Proceedings of the Second Learning on Graphs Conference. Proceedings of Machine Learning Research, vol. 231-20, pp. 1–25 (2024). https://proceedings.mlr.press/v231/yu24a.html

[43] Hijazi, R.H., Jernigan, R.W.: Modelling compositional data using dirichlet regression models. Journal of Applied Probability & Statistics **4**(1), 77–91 (2009)

[44] Tang, Z.-Z., Chen, G.: Zero-inflated generalized dirichlet multinomial regression model for microbiome compositional data analysis. Biostatistics **20**(4), 698–713 (2019) https://doi.org/10.1093/biostatistics/kxy025

[45] Wadsworth, W.D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S.A., Vannucci, M.: An integrative bayesian dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. BMC Bioinformatics **18**(1), 94 (2017) https://doi.org/10.1186/s12859-017-1516-0

[46] Aitchison, J., Shen, S.M.: Logistic-normal distributions: Some properties and uses. Biometrika **67**(2), 261–272 (1980) https://doi.org/10.2307/2335470

[47] Scealy, J.L., Hingee, K.L., Kent, J.T., Wood, A.T.A.: Robust score matching for compositional data. Stat. Comput. **34**(2), 93 (2024) https://doi.org/10.1007/s11222-024-10412-w

[48] Rosenkrantz, R.D.: Where do we stand on maximum entropy? (1978). In: E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics, pp. 210–314. Springer, Dordrecht (1989). https://doi.org/10.1007/978-94-009-6581-2_10

[49] Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., Bonneau, R.A.: Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput. Biol. **11**(5), 1004226 (2015) https://doi.org/10.1371/journal.pcbi.1004226

[50] Jin, S., Notredame, C., Erb, I.: Compositional covariance shrinkage and regularised partial correlations. arXiv [stat.ME] (2022) arXiv:2212.00496 [stat.ME]

[51] Bayes, T.: An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, FRS communicated by mr. price, in a letter to john canton, AMFR S. Philosophical transactions of the Royal Society of London (53), 370–418 (1763)

[52] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data Analysis. Chapman and Hall/CRC, London, UK (2013). https://doi.org/10.1201/b16018 . https://www.taylorfrancis.com/books/9781439898208

[53] Kruschke, J.K.: Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. Academic Press, Boston, MA (2015). https://doi.org/10.1016/B978-0-12-405888-0.09999-2 . https://play.google.com/store/books/details?id=CsOtoAEACAAJ

[54] Murphy, K.P.: Machine Learning: a Probabilistic Perspective. MIT press, Boston, MA (2012). https://play.google.com/store/books/details?id=RC43AgAAQBAJ

[55] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of State Calculations by Fast Computing Machines. J. Chem. Phys. **21**(6), 1087–1092 (1953) https://doi.org/10.1063/1.1699114

[56] Hastings, W.K.: Monte Carlo Sampling Methods using Markov Chains and their Applications. Biometrika **57**(1), 97–109 (1970) https://doi.org/10.1093/biomet/57.1.97

[57] Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D.: Hybrid monte carlo. Phys. Lett. B **195**(2), 216–222 (1987) https://doi.org/10.1016/0370-2693(87)91197-x

[58] Neal, R.M.: MCMC using hamiltonian dynamics. arXiv [stat.CO] (2012) arXiv:1206.1901 [stat.CO]

[59] Betancourt, M.: A conceptual introduction to hamiltonian monte carlo. arXiv [stat.ME] (2017) arXiv:1701.02434 [stat.ME]

[60] Hoffman, M.D., Gelman, A.: The no-U-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. J. Mach. Learn. Res. (2014)

[61] Hyvärinen, A.: Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research **6**, 695–709 (2005)

[62] Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951) https://doi.org/10.1214/aoms/1177729694

[63] Hyvärinen, A.: Some extensions of score matching. Comput. Stat. Data Anal. **51**(5), 2499–2512 (2007) https://doi.org/10.1016/j.csda.2006.09.003

[64] Yu, S., Drton, M., Shojaie, A.: Generalized score matching for non-negative data. J. Mach. Learn. Res. **20** (2019)

[65] Mardia, K.V., Kent, J.T., Laha, A.K.: Score matching estimators for directional distributions. arXiv [math.ST] (2016) arXiv:1604.08470 [math.ST]

[66] Yu, S., Drton, M., Shojaie, A.: Generalized score matching for general domains. Inf inference **11**(2), 739–780 (2022) https://doi.org/10.1093/imaiai/iaaa041

[67] Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., Heath, A.C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J.G., Lozupone, C.A., Lauber, C., Clemente, J.C., Knights, D., Knight, R., Gordon, J.I.: Human gut microbiome viewed across age and geography. Nature **486**(7402), 222–227 (2012) https://doi.org/10.1038/nature11053

[68] Ramirez, K.S., Leff, J.W., Barberán, A., Bates, S.T., Betley, J., Crowther, T.W., Kelly, E.F., Oldfield, E.E., Shaw, E.A., Steenbock, C., Bradford, M.A., Wall, D.H., Fierer, N.: Biogeographic patterns in below-ground diversity in new york city's central park are similar to those observed globally. Proc. Biol. Sci. **281**(1795), 20141988 (2014) https://doi.org/10.1098/rspb.2014.1988

[69] McDonald, D., Hyde, E., Debelius, J.W., Morton, J.T., Gonzalez, A., Ackermann, G., Aksenov, A.A., Behsaz, B., Brennan, C., Chen, Y., DeRight Goldasich, L., Dorrestein, P.C., Dunn, R.R., Fahimipour, A.K., Gaffney, J., Gilbert, J.A., Gogul, G., Green, J.L., Hugenholtz, P., Humphrey, G., Huttenhower, C., Jackson, M.A., Janssen, S., Jeste, D.V., Jiang, L., Kelley, S.T., Knights, D., Kosciolek, T., Ladau, J., Leach, J., Marotz, C., Meleshko, D., Melnik, A.V., Metcalf, J.L., Mohimani, H., Montassier, E., Navas-Molina, J., Nguyen, T.T., Peddada, S., Pevzner, P., Pollard, K.S., Rahnavard, G., Robbins-Pianka, A., Sangwan, N., Shorenstein, J., Smarr, L., Song, S.J., Spector, T., Swafford, A.D., Thackray, V.G., Thompson, L.R., Tripathi, A., Vázquez-Baeza, Y., Vrbanac, A., Wischmeyer, P., Wolfe, E., Zhu, Q., American Gut Consortium, Knight, R.: American gut: An open platform for citizen science microbiome research. mSystems **3**(3) (2018) https://doi.org/10.1128/mSystems.00031-18

[70] Human Microbiome Project Consortium: A framework for human microbiome research. Nature **486**(7402), 215–221 (2012) https://doi.org/10.1038/nature11209

[71] Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J.C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C.P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T.N., Shalek, A., Shapiro, E., Sharma, P., Shin, J.W., Stegle, O., Stratton, M., Stubbington, M.J.T., Theis, F.J., Uhlen, M., Oudenaarden, A., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., Yosef, N., Human Cell Atlas Meeting Participants: The human cell atlas. Elife **6** (2017) https://doi.org/10.7554/eLife.27041

[72] Smillie, C.S., Biton, M., Ordovas-Montanes, J., Sullivan, K.M., Burgin, G., Graham, D.B., Herbst, R.H., Rogel, N., Slyper, M., Waldman, J., Sud, M., Andrews, E., Velonias, G., Haber, A.L., Jagadeesh, K., Vickovic, S., Yao, J., Stevens, C., Dionne, D., Nguyen, L.T., Villani, A.-C., Hofree, M., Creasey, E.A., Huang, H., Rozenblatt-Rosen, O., Garber, J.J., Khalili, H., Desch, A.N., Daly, M.J., Ananthakrishnan, A.N., Shalek, A.K., Xavier, R.J., Regev, A.: Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. Cell **178**(3), 714–73022 (2019) https://doi.org/10.1016/j.cell.2019.06.029

[73] Haber, A.L., Biton, M., Rogel, N., Herbst, R.H., Shekhar, K., Smillie, C., Burgin, G., Delorey, T.M., Howitt, M.R., Katz, Y., Tirosh, I., Beyaz, S., Dionne, D., Zhang, M., Raychowdhury, R., Garrett, W.S., Rozenblatt-Rosen, O., Shi, H.N., Yilmaz, O., Xavier, R.J., Regev, A.: A single-cell survey of the small intestinal epithelium. Nature **551**(7680), 333–339 (2017) https://doi.org/10.1038/nature24489

[74] Perez, R.K., Gordon, M.G., Subramaniam, M., Kim, M.C., Hartoularos, G.C., Targ, S., Sun, Y., Ogorodnikov, A., Bueno, R., Lu, A., Thompson, M., Rappoport, N., Dahl, A., Lanata, C.M., Matloubian, M., Maliskova, L., Kwek, S.S., Li, T., Slyper, M., Waldman, J., Dionne, D., Rozenblatt-Rosen, O., Fong, L., Dall'Era, M., Balliu, B., Regev, A., Yazdany, J., Criswell, L.A., Zaitlen, N., Ye, C.J.: Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. Science **376**(6589), 1970 (2022) https://doi.org/10.1126/science.abf1970

[75] Carcy, S., Ostner, J., Tran, V., Menden, M., Müller, C.L.: MetaIBS - large-scale amplicon-based meta analysis of irritable bowel syndrome. bioRxiv, 2024–0122575775 (2024) https://doi.org/10.1101/2024.01.22.575775

[76] Tsoucas, D., Dong, R., Chen, H., Zhu, Q., Guo, G., Yuan, G.-C.: Accurate estimation of cell-type composition from gene expression data. Nat. Commun. **10**(1), 2975 (2019) https://doi.org/10.1038/s41467-019-10802-z

[77] Hashimoto, K., Kouno, T., Ikawa, T., Hayatsu, N., Miyajima, Y., Yabukami, H., Terooatea, T., Sasaki, T., Suzuki, T., Valentine, M., Pascarella, G., Okazaki, Y., Suzuki, H., Shin, J.W., Minoda, A., Taniuchi, I., Okano, H., Arai, Y., Hirose, N., Carninci, P.: Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians. Proc. Natl. Acad. Sci. U. S. A. **116**(48), 24242–24251 (2019) https://doi.org/10.1073/pnas.1907883116

[78] Betancourt, M., Girolami, M.: Hamiltonian monte carlo for hierarchical models. In: Current Trends in Bayesian Methodology with Applications, pp. 79–101. Chapman and Hall/CRC, London, UK (2015). https://doi.org/10.1201/b18502-5 . http://www.crcnetbase.com/doi/10.1201/b18502-5

[79] Friedman, N.: Inferring cellular networks using probabilistic graphical models. Science **303**(5659), 799–805 (2004) https://doi.org/10.1126/science.1094068

[80] Bonneau, R.: Learning biological networks: from modules to dynamics. Nat. Chem. Biol. **4**(11), 658–664 (2008) https://doi.org/10.1038/nchembio.122

[81] Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Series B Stat. Methodol. **58**(1), 267–288 (1996)

[82] Wu, Y., Wang, L.: A survey of tuning parameter selection for high-dimensional regression. arXiv [stat.ME] (2019) arXiv:1908.03669 [stat.ME]

[83] Akaike, H.: A new look at the statistical model identification. IEEE Trans. Automat. Contr. **19**(6), 716–723 (1974) https://doi.org/10.1109/tac.1974.1100705

[84] Chen, J., Chen, Z.: Extended bayesian information criteria for model selection with large model spaces. Biometrika **95**(3), 759–771 (2008) https://doi.org/10.1093/biomet/asn034

[85] Foygel, R., Drton, M.: Extended bayesian information criteria for gaussian graphical models. arXiv [math.ST] (2010) arXiv:1011.6640 [math.ST]

[86] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, 2nd edn. Springer series in statistics. Springer, New York, NY (2009). https://doi.org/10.1007/978-0-387-84858-7 . https://link.springer.com/book/10.1007/978-0-387-84858-7

[87] Chetverikov, D., Liao, Z., Chernozhukov, V.: On cross-validated lasso in high dimensions. arXiv [math.ST] (2016) arXiv:1605.02214 [math.ST]

[88] Lin, L., Drton, M., Shojaie, A.: Estimation of high-dimensional graphical models using regularized score matching. Electron. J. Stat. **10**(1), 806–854 (2016) https://doi.org/10.1214/16-EJS1126

[89] Mitchell, T.J., Beauchamp, J.J.: Bayesian variable selection in linear regression. J. Am. Stat. Assoc. **83**(404), 1023–1032 (1988) https://doi.org/10.1080/01621459.1988.10478694

[90] George, E., McCulloch, R.: Approaches for bayesian variable selection. Statistica Sinica **7**(2), 339–373 (1997)

[91] Ishwaran, H., Rao, J.S.: Spike and slab variable selection: Frequentist and bayesian strategies. Ann. Stat. **33**(2), 730–773 (2005)

[92] Malsiner-Walli, G., Wagner, H.: Comparing spike and slab priors for bayesian variable selection. arXiv [stat.ME] (2018) arXiv:1812.07259 [stat.ME]

[93] Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**(6), 721–741 (1984) https://doi.org/10.1109/tpami.1984.4767596

[94] Thomson, W., Jabbari, S., Taylor, A.E., Arlt, W., Smith, D.J.: Simultaneous parameter estimation and variable selection via the logit-normal continuous analogue of the spike-and-slab prior. J. R. Soc. Interface **16**(150), 20180572 (2019) https://doi.org/10.1098/rsif.2018.0572

[95] Ročková, V., George, E.I.: The spike-and-slab LASSO. J. Am. Stat. Assoc. **113**(521), 431–444 (2018) https://doi.org/10.1080/01621459.2016.1260469

[96] Bai, R., Rockova, V., George, E.I.: Spike-and-slab meets LASSO: A review of the spike-and-slab LASSO. arXiv [stat.ME] (2020) arXiv:2010.06451 [stat.ME]

[97] Yang, L., Chen, J.: A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions. Microbiome **10**(1), 130 (2022) https://doi.org/10.1186/s40168-022-01320-0

[98] Fernandes, A.D., Reid, J.N., Macklaim, J.M., McMurrough, T.A., Edgell, D.R., Gloor, G.B.: Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome **2**, 15 (2014) https://doi.org/10.1186/2049-2618-2-15

[99] Hawinkel, S., Mattiello, F., Bijnens, L., Thas, O.: A broken promise: microbiome differential abundance methods do not control the false discovery rate. Brief. Bioinform. **20**(1), 210–221 (2019) https://doi.org/10.1093/bib/bbx104

[100] Nearing, J.T., Douglas, G.M., Hayes, M.G., MacDonald, J., Desai, D.K., Allward, N., Jones, C.M.A., Wright, R.J., Dhanani, A.S., Comeau, A.M., Langille, M.G.I.: Microbiome differential abundance methods produce different results across 38 datasets. Nat. Commun. **13**(1), 342 (2022) https://doi.org/10.1038/s41467-022-28034-z

[101] Calgaro, M., Romualdi, C., Waldron, L., Risso, D., Vitulo, N.: Assessment of statistical methods from single cell, bulk RNA-seq and metagenomics applied to microbiome data. bioRxiv, 2020–0115907964 (2020) https://doi.org/10.1101/2020.01.15.907964

[102] Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J.R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E.R., Knight, R.: Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome **5**(1), 27 (2017) https://doi.org/10.1186/s40168-017-0237-y

[103] Wang, S.: Robust differential abundance test in compositional data. Biometrika **110**(1), 169–185 (2023) https://doi.org/10.1093/biomet/asac029

[104] Brill, B., Amir, A., Heller, R.: Testing for differential abundance in compositional counts data, with application to microbiome studies. arXiv [q-bio.GN] (2019) arXiv:1904.08937 [q-bio.GN]

[105] Mandal, S., Van Treuren, W., White, R.A., Eggesbø, M., Knight, R., Peddada, S.D.: Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb. Ecol. Health Dis. **26**, 27663 (2015) https://doi.org/10.3402/mehd.v26.27663

[106] Kaul, A., Mandal, S., Davidov, O., Peddada, S.D.: Analysis of microbiome data in the presence of excess zeros. Front. Microbiol. **8**, 2114 (2017) https://doi.org/10.3389/fmicb.2017.02114

[107] Lin, H., Peddada, S.D.: Analysis of compositions of microbiomes with bias correction. Nat. Commun. **11**(1), 3514 (2020) https://doi.org/10.1038/s41467-020-17041-7

[108] Lin, H., Peddada, S.D.: Multigroup analysis of compositions of microbiomes with covariate adjustments and repeated measures. Nat. Methods **21**(1), 83–91 (2024) https://doi.org/10.1038/s41592-023-02092-7

[109] Zhou, H., He, K., Chen, J., Zhang, X.: LinDA: linear models for differential abundance analysis of microbiome compositional data. Genome Biol. **23**(1), 95 (2022) https://doi.org/10.1186/s13059-022-02655-5 arXiv:2104.00242 [stat.ME]

[110] Mi, K., Xu, Y., Li, Y., Liu, X.: QMD: A new method to quantify microbial absolute abundance differences between groups. Imeta (2023) https://doi.org/10.1002/imt2.78

[111] Lin, X., Chau, C., Huang, Y., Ho, J.W.K.: DCATS: differential composition analysis for complex single-cell experimental designs. bioRxiv, 2022–0321485232 (2022) https://doi.org/10.1101/2022.03.21.485232

[112] Ma, S., Huttenhower, C., Janson, L.: Compositional differential abundance testing: Defining and finding a new type of health-microbiome associations. bioRxiv, 2024–0604596112 (2024) https://doi.org/10.1101/2024.06.04.596112

[113] Li, Z., Lee, K., Karagas, M.R., Madan, J.C., Hoen, A.G., O'Malley, A.J., Li, H.: Conditional regression based on a multivariate zero-inflated logistic normal model for microbiome relative abundance data. arXiv [stat.AP], 1709–07798 (2017) https://doi.org/10.48550/ARXIV.1709.07798 1709.07798 [stat.AP]

[114] Jiang, S., Xiao, G., Koh, A.Y., Kim, J., Li, Q., Zhan, X.: A bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data. Biostatistics **22**(3), 522–540 (2021) https://doi.org/10.1093/biostatistics/kxz050

[115] Hu, Y., Satten, G.A., Hu, Y.-J.: LOCOM: A logistic regression model for testing differential abundance in compositional microbiome data with false discovery rate control. bioRxiv, 2021–1003462964 (2021) https://doi.org/10.1101/2021.10.03.462964

[116] Martin, B.D., Witten, D., Willis, A.D.: Modeling microbial abundances and dysbiosis with beta-binomial regression. arXiv [stat.ME] (1), 94–115 (2019) https://doi.org/10.1214/19-aoas1283 arXiv:1902.02776 [stat.ME]

[117] Dann, E., Henderson, N.C., Teichmann, S.A., Morgan, M.D., Marioni, J.C.: Differential abundance testing on single-cell data using k-nearest neighbor graphs. Nat. Biotechnol., 1–9 (2021) https://doi.org/10.1038/s41587-021-01033-z

[118] Zhou, C., Zhao, H., Wang, T.: Transformation and differential abundance analysis of microbiome data incorporating phylogeny. Bioinformatics (2021) https://doi.org/10.1093/bioinformatics/btab543

[119] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. **57**(1), 289–300 (1995)

[120] Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. Ann. Stat. **29**(4), 1165–1188 (2001) https://doi.org/10.1214/aos/1013699998

[121] Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics **6**(2), 65–70 (1979) https://doi.org/10.2307/4615733

[122] Muller, P., Parmigiani, G., Rice, K.: FDR and bayesian multiple comparisons rules. Johns Hopkins University, Dept. of Biostatistics Working Papers. **115**(115), 1–15 (2006)

[123] Scott, J.G., Berger, J.O.: Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. aos **38**(5), 2587–2619 (2010) https://doi.org/10.1214/10-AOS792

[124] Newton, M.A., Noueiry, A., Sarkar, D., Ahlquist, P.: Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics **5**(2), 155–176 (2004) https://doi.org/10.1093/biostatistics/5.2.155

[125] Harrison, J.G., Calder, W.J., Shastry, V., Buerkle, C.A.: Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. Mol. Ecol. Resour. **20**(2), 481–497 (2020) https://doi.org/10.1111/1755-0998.13128

[126] Egozcue, J.J., Graffelman, J., Ortego, M.I., Pawlowsky-Glahn, V.: Some thoughts on counts in sequencing studies. NAR Genom Bioinform **2**(4), 094 (2020) https://doi.org/10.1093/nargab/lqaa094

[127] Xia, F., Chen, J., Fung, W.K., Li, H.: A logistic normal multinomial regression model for microbiome compositional data analysis. Biometrics **69**(4), 1053–1063 (2013) https://doi.org/10.1111/biom.12079

[128] Lubbe, S., Filzmoser, P., Templ, M.: Comparison of zero replacement strategies for compositional data with large numbers of zeros. Chemometrics Intellig. Lab. Syst. **210**, 104248 (2021) https://doi.org/10.1016/j.chemolab.2021.104248

[129] Ahlmann-Eltze, C., Huber, W.: Comparison of transformations for single-cell RNA-seq data. Nat. Methods (2023) https://doi.org/10.1038/s41592-023-01814-1

[130] Oehlert, G.W.: A note on the delta method. Am. Stat. **46**(1), 27 (1992) https://doi.org/10.2307/2684406

[131] Box, G.E.P., Cox, D.R.: An analysis of transformations. J. R. Stat. Soc. Series B Stat. Methodol. **26**(2), 211–243 (1964) https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

[132] Tsagris, M., Preston, S., Wood, A.T.A.: Improved classification for compositional data using the α-transformation. J. Classif. **33**(2), 243–261 (2016) https://doi.org/10.1007/s00357-016-9207-5

[133] Greenacre, M.: The chiPower transformation: a valid alternative to logratio transformations in compositional data analysis. Adv. Data Anal. Classif. (2024) https://doi.org/10.1007/s11634-024-00600-x

[134] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O.: The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. **41**(Database issue), 590–6 (2013) https://doi.org/10.1093/nar/gks1219

[135] McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., Hugenholtz, P.: An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. **6**(3), 610–618 (2012) https://doi.org/10.1038/ismej.2011.139

[136] Schliep, K.P.: phangorn: phylogenetic analysis in R. Bioinformatics **27**(4), 592–593 (2010) https://doi.org/10.1093/bioinformatics/btq706

[137] Morelli, L., Giansanti, V., Cittaro, D.: Nested stochastic block models applied to the analysis of single cell data. BMC Bioinformatics **22**(1), 576 (2021) https://doi.org/10.1186/s12859-021-04489-7

[138] Yan, X., Bien, J.: Rare feature selection in high dimensions. J. Am. Stat. Assoc. **116**(534), 887–900 (2021) https://doi.org/10.1080/01621459.2020.1796677

[139] Bien, J., Yan, X., Simpson, L., Müller, C.L.: Tree-aggregated predictive modeling of microbiome data. Sci. Rep. **11**(1), 14505 (2021) https://doi.org/10.1038/s41598-021-93645-3

[140] Neufeld, A., Dharamshi, A., Gao, L.L., Witten, D.: Data thinning for convolution-closed distributions. arXiv [stat.ME] (2023) arXiv:2301.07276 [stat.ME]

[141] Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. **2008**(10), 10008 (2008) https://doi.org/10.1088/1742-5468/2008/10/P10008

[142] Traag, V., Waltman, L., Eck, N.J.: From louvain to leiden: guaranteeing well-connected communities. Sci. Rep. **9** (2018) https://doi.org/10.1038/s41598-019-41695-z 1810.08473

[143] Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., Lindauer, M.: Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. (2023) https://doi.org/10.1002/widm.1484

[144] Fu, W., Perry, P.O.: Estimating the number of clusters using cross-validation. J. Comput. Graph. Stat. **29**(1), 162–173 (2020) https://doi.org/10.1080/10618600.2019.1647846

[145] Owen, A., Perry, P.O.: Bi-cross-validation of the SVD and the nonnegative matrix factorization. The Annals of Applied Statistics **3**(2), 564–594 (2009) https://doi.org/10.1214/08-AOAS227 0908.2062

[146] Neufeld, A., Popp, J., Gao, L.L., Battle, A., Witten, D.: Negative binomial count splitting for single-cell RNA sequencing data. arXiv [stat.ME] (2023) arXiv:2307.12985 [stat.ME]

[147] Jørgensen, B., Peter, X.: Stationary time series models with exponential dispersion model margins. J. Appl. Probab. **35**(1), 78–92 (1998) https://doi.org/10.1239/jap/1032192553

[148] Xia, L., Lee, C., Li, J.J.: Statistical method scDEED for detecting dubious 2D single-cell embeddings and optimizing t-SNE and UMAP hyperparameters. Nat. Commun. **15**(1), 1753 (2024) https://doi.org/10.1038/s41467-024-45891-y

[149] Li, J.J.: Using Synthetic Null Data to Enhance Statistical Rigor in Genomics. ISMB/ECCB 2023 keynote presentation (2023). http://jsb.ucla.edu/sites/default/files/072523_Overton.pdf

[150] Breugel, B., Liu, T., Oglic, D., Schaar, M.: Synthetic data in biomedicine via generative artificial intelligence. Nat Rev Bioeng, 1–14 (2024) https://doi.org/10.1038/s44222-024-00245-7

[151] Fisher, R.A.: The Design of Experiments. Oliver and Boyd, Edniburgh, UK (1935)

[152] Song, D., Li, K., Ge, X., Li, J.J.: ClusterDE: a post-clustering differential expression (DE) method robust to false-positive inflation caused by double dipping. bioRxiv (2023) https://doi.org/10.1101/2023.07.21.550107

[153] Nelsen, R.B.: An Introduction to Copulas, 2nd edn. Springer Series in Statistics. Springer, New York, NY (2006). https://doi.org/10.1007/0-387-28678-0 . https://link.springer.com/book/10.1007/0-387-28678-0

[154] Cario, M.C., Nelson, B.L.: Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical Report 1, Technical Report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois (1997). http://www.ressources-actuarielles.net/EXT/ISFA/1226.nsf/769998e0a65ea348c1257052003eb94f/5d499a3efc8ae4dfc125756c00391ca6/$FILE/NORTA.pdf

[155] Li, W.V., Li, J.J.: A statistical simulator scDesign for rational scRNA-seq experimental design. Bioinformatics **35**(14), 41–50 (2019) https://doi.org/10.1093/bioinformatics/btz321

[156] Sun, T., Song, D., Li, W.V., Li, J.J.: scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. Genome Biol. **22**(1), 163 (2021) https://doi.org/10.1186/s13059-021-02367-2

[157] Song, D., Wang, Q., Yan, G., Liu, T., Sun, T., Li, J.J.: scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics. Nat. Biotechnol. (2023) https://doi.org/10.1038/s41587-023-01772-1

[158] Rüschendorf, L.: Copulas, sklar's theorem, and distributional transform. In: Springer Series in Operations Research and Financial Engineering. Springer Series in Operations Research and Financial Engineering, pp. 3–34. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-33590-7_1

[159] Neufeld, A., Gao, L.L., Popp, J., Battle, A., Witten, D.: Inference after latent variable estimation for single-cell RNA sequencing data. arXiv [stat.ME] (2022) arXiv:2207.00554 [stat.ME]

[160] Zhang, J.M., Kamath, G.M., Tse, D.N.: Valid post-clustering differential analysis for single-cell RNA-seq. Cell Systems **9**(4), 383–3926 (2019) https://doi.org/10.1016/j.cels.2019.07.012

[161] Ge, X., Chen, Y.E., Song, D., McDermott, M., Woyshner, K., Manousopoulou, A., Wang, N., Li, W., Wang, L.D., Li, J.J.: Clipper: p-value-free FDR control on high-throughput data from two conditions. Genome Biol. **22**(1), 288 (2021) https://doi.org/10.1186/s13059-021-02506-9

[162] Heumos, L., Ji, Y., May, L., Green, T., Zhang, X., Wu, X., Ostner, J., Peidli, S., Schumacher, A., Hrovatin, K., Müller, M., Chong, F., Sturm, G., Tejada, A., Dann, E., Dong, M., Bahrami, M., Gold, I., Rybakov, S., Namsaraeva, A., Moinfar, A., Zheng, Z., Roellin, E., Mekki, I., Sander, C., Lotfollahi, M., Schiller, H.B., Theis, F.J.: Pertpy: an end-to-end framework for perturbation analysis. bioRxiv, 2024–0804606516 (2024) https://doi.org/10.1101/2024.08.04.606516

[163] Labus, J.S., Hollister, E.B., Jacobs, J., Kirbach, K., Oezguen, N., Gupta, A., Acosta, J., Luna, R.A., Aagaard, K., Versalovic, J., Savidge, T., Hsiao, E., Tillisch, K., Mayer, E.A.: Differences in gut microbial composition correlate with regional brain volumes in irritable bowel syndrome. Microbiome **5**(1), 49 (2017) https://doi.org/10.1186/s40168-017-0260-z

[164] Fernandez, L., Rosvall, M., Normark, J., Fällman, M., Avican, K.: Co-PATHOgenex web application for assessing complex stress responses in pathogenic bacteria. Microbiol Spectr **12**(1), 0278123 (2024) https://doi.org/10.1128/spectrum.02781-23

[165] Chen, E.Z., Li, H.: A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics **32**(17), 2611–2617 (2016) https://doi.org/10.1093/bioinformatics/btw308

[166] Äijö, T., Müller, C.L., Bonneau, R.: Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing. Bioinformatics **34**(3), 372–380 (2018) https://doi.org/10.1093/bioinformatics/btx549

[167] Nelder, J.A., Wedderburn, R.W.M.: Generalized linear models. J. R. Stat. Soc. Ser. A **135**(3), 370 (1972) https://doi.org/10.2307/2344614

[168] Hastie, T., Tibshirani, R.: Generalized additive models. Stat. Sci. **1**(3), 297–310 (1986) https://doi.org/10.1214/ss/1177013604

[169] Kassel, L.H., Scealy, J., Bell, B.M.: Score Matching Estimation by Automatic Differentiation (2024). https://cloud.r-project.org/web/packages/scorematchingad/scorematchingad.pdf

[170] Rügamer, D., Kolb, C., Fritz, C., Pfisterer, F., Kopper, P., Bischl, B., Shen, R., Bukas, C., Sousa, L.B.d.A.e., Thalmeier, D., Baumann, P., Kook, L., Klein, N., Müller, C.L.: Deepregression: A flexible neural network framework for semi-structured deep distributional regression. arXiv [stat.ML] (2021) arXiv:2104.02705 [stat.ML]

[171] Rügamer, D., Kolb, C., Klein, N.: Semi-structured distributional regression. Am. Stat. **78**(1), 88–99 (2024) https://doi.org/10.1080/00031305.2022.2164054

[172] Zappia, L., Phipson, B., Oshlack, A.: Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLoS Comput. Biol. **14**(6), 1–14 (2018) https://doi.org/10.1371/journal.pcbi.1006245

[173] Hasin, Y., Seldin, M., Lusis, A.: Multi-omics approaches to disease. Genome Biol. **18**(1), 83 (2017) https://doi.org/10.1186/s13059-017-1215-1

A. Bayesian modeling of compositional HTS data

A.1. scCODA is a Bayesian model for compositional single-cell data analysis

Contributing article

Büttner, M.*, **Ostner, J.***, Müller, C. L., Theis, F. J., and Schubert, B. (2021). scCODA is a Bayesian model for compositional single-cell data analysis. *Nat. Commun. 12, 6876.* doi: `https://doi.org/10.1038/s41467-021-27150-6`
* joint first co-authorship

Replication code

Source code for this contribution has been deposited on Github at `https://github.com/theislab/sccoda`. All code to reproduce the presented analyses can be found on Github at `https://github.com/theislab/scCODA_reproducibility`.

Copyright information

Author contributions

M.B. and B.S. conceived of the study. J.O. developed scCODA and conducted the benchmarking study. M.B., J.O., and B.S. analyzed the data. C.L.M. helped design the model comparison. B.S. and F.J.T. supervised the study and model development. B.S. designed the benchmarking study and conducted the power analysis. M.B., J.O., C.L.M., and B.S. wrote the manuscript. All authors read and approved the final manuscript.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-27150-6 OPEN

scCODA is a Bayesian model for compositional single-cell data analysis

M. Büttner [1,6], J. Ostner [1,2,6], C. L. Müller [1,2,3✉], F. J. Theis [1,4,5,7] & B. Schubert [1,4,7✉]

Compositional changes of cell types are main drivers of biological processes. Their detection through single-cell experiments is difficult due to the compositionality of the data and low sample sizes. We introduce scCODA (https://github.com/theislab/scCODA), a Bayesian model addressing these issues enabling the study of complex cell type effects in disease, and other stimuli. scCODA demonstrated excellent detection performance, while reliably controlling for false discoveries, and identified experimentally verified cell type changes that were missed in original analyses.

[1] Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [2] Department of Statistics, Ludwig-Maximilians-Universität München, München, Germany. [3] Center for Computational Mathematics, Flatiron Institute, New York, NY, USA. [4] Department of Mathematics, Technische Universität München, Garching bei München, Germany. [5] TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany. [6] These authors contributed equally: M. Büttner, J. Ostner. [7] These authors jointly supervised this work: F. J. Theis, B. Schubert. ✉email: christian.mueller@stat.uni-muenchen.de; benjamin.schubert@helmholtz-muenchen.de

Recent advances in single-cell RNA-sequencing (scRNA-seq) allow large-scale quantitative transcriptional profiling of individual cells across a wide range of tissues, thus enabling the monitoring of transcriptional changes between conditions or developmental stages and the data-driven identification of distinct cell types.

Although being important drivers of biological processes such as in disease[1], development[2], aging[3], and immunity[4], shifts in cell-type compositions are non-trivial to detect using scRNA-seq. Statistical tests need to account for multiple sources of technical and methodological limitations, including the low number of experimental replications. The total number of cells per sample is restricted in most single-cell technologies, implying that cell-type counts are proportional in nature. This, in turn, leads to a negative bias in cell-type correlation estimation[5] (Fig. 1a). For example, if only a specific cell type is depleted after perturbation, the relative frequency of others will rise. If taken at face value, this would lead to an inflation of differential cell types. Therefore, standard univariate statistical models that test compositional changes of each cell type independently may falsely deem certain population shifts as real effects, even though they were solely induced by the inherent negative correlations of the cell-type proportions (Fig. 1b). Yet, common statistical approaches currently applied in compositional cell-type analysis ignore this effect. For example, Haber et al.[6] applied a univariate test based on Poisson regression, Hashimoto et al.[3] a Wilcoxon rank-sum test, and Cao et al.[7] proposed a method based on a generalized linear regression framework with a Poisson likelihood, all thus not addressing the issue of compositionality.

To account for the inherent bias present in cell-type compositions, we drew inspiration from methods for compositional analysis of microbiome data[8,9] and propose a Bayesian approach for cell-type composition differential abundance analysis to further address the low replicate issue. The single-cell compositional data analysis (scCODA) framework models cell-type counts with a hierarchical Dirichlet-Multinomial distribution that accounts for the uncertainty in cell-type proportions and the negative correlative bias via joint modeling of all measured cell-type proportions instead of individual ones (Fig. 1c, Methods—"Model description"). The model uses a Logit-normal spike-and-slab prior[10] with a log-link function to estimate effects of binary (or continuous) covariates on cell-type proportions in a parsimonious fashion. Since compositional analysis always requires a reference to be able to identify compositional changes[5], scCODA can automatically select an appropriate cell type as the reference (Methods—"Automatic reference selection") or uses a prespecified reference cell type[11]. This implies that credible changes detected by scCODA have to be interpreted in relation to the selected reference. On top, the framework offers access to other well-established compositional test statistics and is fully integrated into the Scanpy[12] ecosystem.

Results

scCODA performs best in a benchmark of synthetic datasets. We first performed comprehensive benchmarks on synthetic data across a wide range of scenarios (Methods—"Simulation") that focused on scCODA's primary application: the behavior of a single binary covariate that models the effect of a perturbation of interest in the respective scRNA-seq experiment. To detect statistically credible changes in cell-type compositions, we calculate the model inclusion probability for each covariate determined by the spike-and-slab prior (Methods—"Model description"). By using a direct posterior probability approach, scCODA automatically determines a cutoff on the posterior inclusion probability for credible effects that controls for the false discovery rate (FDR, Methods—"Spike-and-slab threshold determination").

We compared scCODA's performance to state-of-the-art differential compositional testing schemes from the microbiome field as well as all non-compositional tests recently applied to

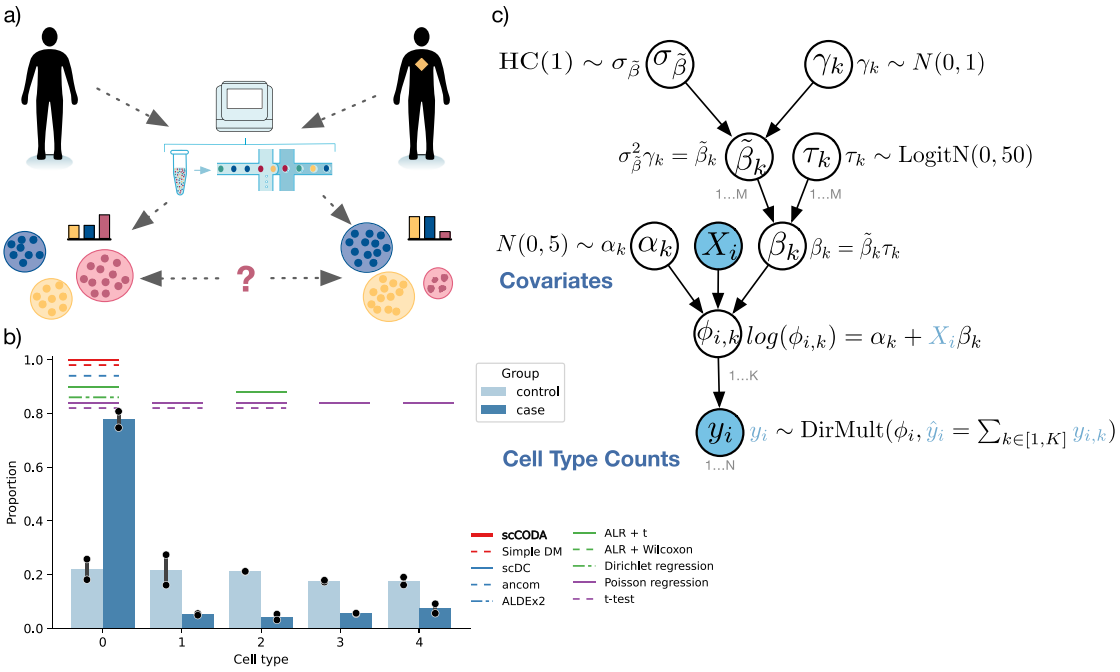


Fig. 1 Compositional data analysis in single-cell RNA-sequencing data. a Single-cell analysis of control and disease states of a human tissue sample. Disease states reflect changes in the cell-type composition. **b** Exemplary realization of the tested scenarios with high compositional log-fold change and low replicate number ($n = 2$ samples per group). Colored horizontal lines indicate statistically detected compositional changes between case and control for different methods. The error bars denote the 95% confidence interval around the mean. **c** The scCODA model structure with hyperparameters. Blue variables are observed. *DirMult* indicates a Dirichlet-Multinomial, *N* a Normal, *logitN* a Logit-Normal, and *HC* a Half-Cauchy distribution.

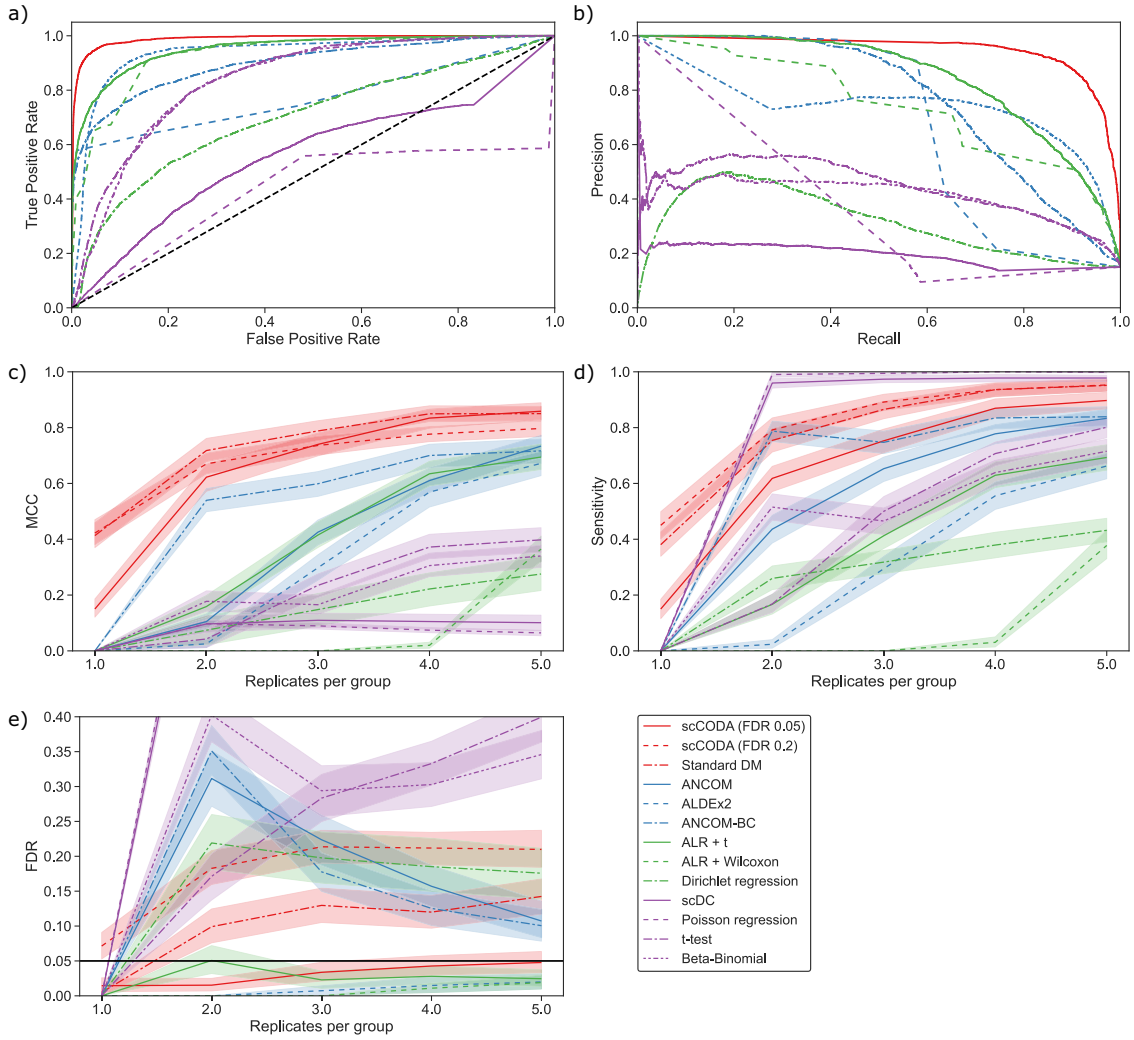


Fig. 2 Comparison of scCODA's benchmark performance to other differential abundance testing methods. Bayesian models (red), non-standard compositional models (blue), compositional tests/regression (green), non-compositional methods (purple). Shaded areas represent 95% confidence intervals. **a** Receiver-operating curve (n >1 samples per group). AUC scores are reported in (Supplementary Table 1)**. b** Precision-recall curve (n >1 samples per group). Average precision scores are reported in (Supplementary Table 1)**. c**–**e** Performance metrics with increasing number of replicates per group over all tested scenarios. In the case of $n = 1$ sample per group, only Bayesian methods are applicable, other methods cannot detect any changes. **c** Overall performance measured by Matthews' correlation coefficient (MCC). **d** Sensitivity measured by true positive rate (TPR). **e** Precision measured by false discovery rate (FDR). The nominal FDR level of 0.05 for all methods (except scCODA with FDR 0.2) is indicated with a horizontal black line.

single-cell data (Fig. 2), all with a nominal FDR level of 0.05. In our synthetic benchmarks, we found scCODA to significantly outperform all non-Bayesian approaches in the regime of low-sample sizes across a wide variety of effects and experimental settings with an average Matthews' correlation coefficient (MCC) of 0.64. Considering the number of replicates per group, the Bayesian models (scCODA and a standard Dirichlet-multinomial modeling approach; red lines in Fig. 2) had a considerable edge over all other methods in the common scenario with a low number of replicates per group, and increased their MCC further with the sample size (Fig. 2c). Other compositional non-Bayesian models such as ANCOM-BC[13], ANCOM[14], ALDEx2[15], and additive log-ratio (ALR) transformed proportions combined with a t test (Methods—"Model comparison") showed similar behavior, albeit with lower MCC. Non-compositional models, such as the Beta-Binomial model[16], the scDC model[7], or univariate t tests, (purple lines in Fig. 2) included more false positives with increasing effect size (Fig. 2d, e) and the number of replicates per group, highlighting the need for a compositional adjustment when modeling population data from scRNA-seq.

Looking at the false discovery rate (Fig. 2d), we could confirm recent findings that ANCOM and ANCOM-BC show increased numbers of false-positive results, especially in the low-sample setting[17,18]. Also, the standard Dirichlet-Multinomial model showed an average false discovery rate at almost twice the nominal level of 0.05. Only scCODA, ALDEx2, and the ALR-transformed statistical tests were able to accurately control for the false discovery rate in all scenarios. Of these methods, scCODA showed the best sensitivity (true positive rate; Fig. 2e) by a large margin. A more detailed look at the results in terms of effect size and the number of cell types is shown in (Supplementary Figs. 1–3).

When increasing the expected FDR level of scCODA to 0.2, the model sensitivity increased at the cost of a higher false discovery rate, which is controlled by the nominal FDR level (Fig. 2c–e).

Since non-Bayesian methods are not able to produce any results for the case of one sample per group due to a lack of degrees of freedom, we assumed no discoveries on these datasets, resulting in MCC, TPR, and FDR of 0. In contrast, Bayesian models adjust prior assumptions by the evidence from the data. Therefore, tests on one-sample data are possible, albeit with a strong influence from the

choice of priors. Because scCODA gives equal prior probability to exclusion and inclusion of an effect (Methods—"Model description"), the selection of credible effects is driven by the data, even when the sample size is small. Supplementary Fig. 3 shows that Bayesian models can still detect some very strong effects (increase = 2000), even in the one-sample case.

We also performed sensitivity analysis by the receiver-operating characteristic and precision-recall curve (Fig. 2a, b and Supplementary Table 1). To allow for a fair comparison of frequentist and Bayesian methods, we only considered the case of more than one sample per group for all methods, since frequentist tests are not applicable in the one-sample case. Furthermore, we excluded the standard Dirichlet-Multinomial model from the comparison due to problematic thresholding. In both metrics, scCODA outperformed all other tested methods (AUC = 0.99; average precision Score=0.94). Most other compositional methods also showed adequate ability to accurately recover the true effects, while non-compositional methods were among the worst-performing methods.

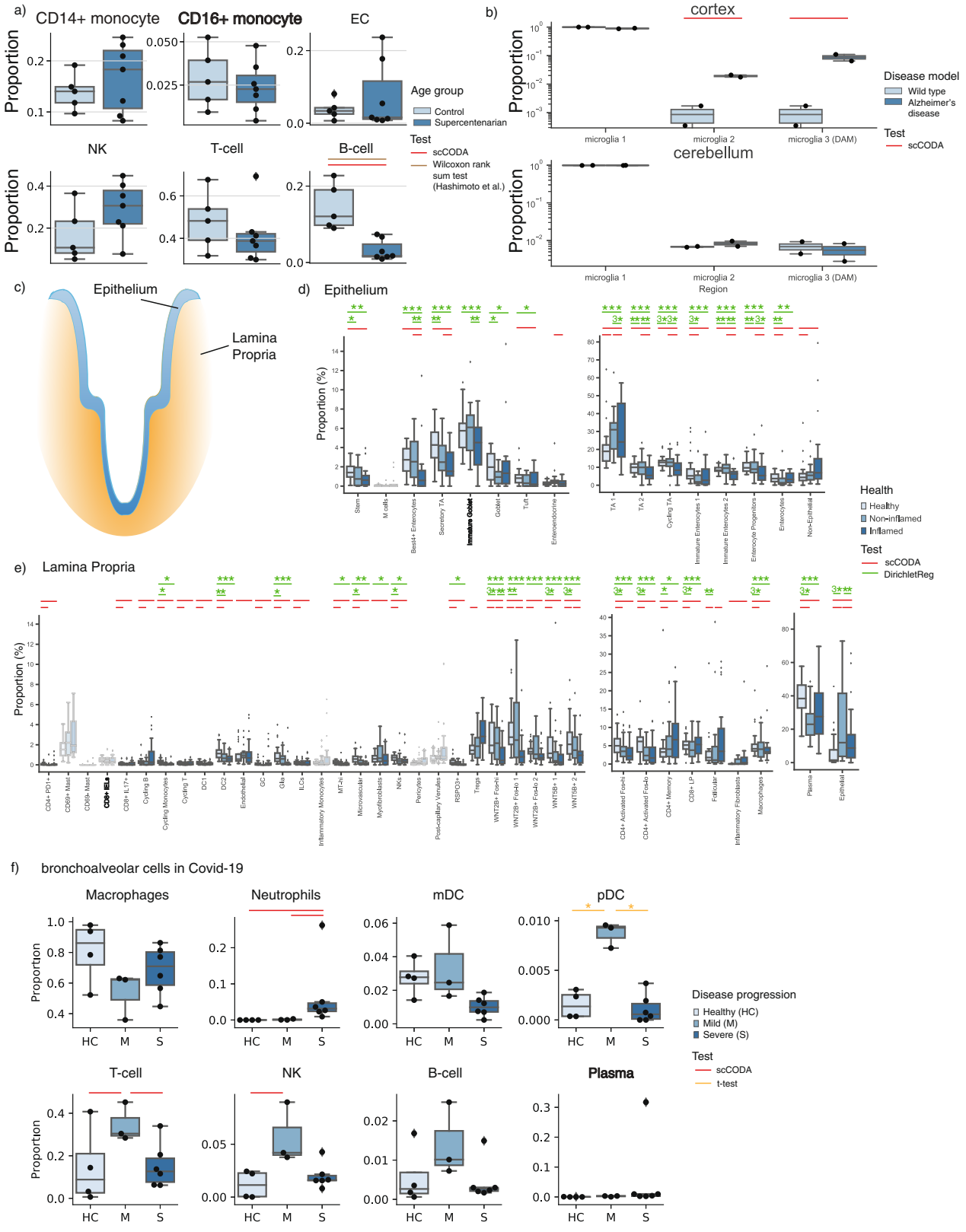
While scCODA performs better than other methods in the low-sample case, we stress that analyses on datasets with larger sample sizes will always be less sensitive to outliers and variability in the data. To determine a reliable sample size for detecting effects of different strengths, we conducted a power analysis of our method.

Power analysis to detect compositional changes. Since extensive replication of scRNA-seq experiments is still costly and hence rare, yet essential for studying compositional changes, we also investigated the sample size dependency of effect size and rarity of affected cell type on scCODA's performance (Supplementary Fig. 4d–f). We performed a power analysis fitting a quasibinomial model (R^2 = [0.937, 0.9377, 0.936] for FDR = [0.05, 0.1, 0.2], Methods—"Power analysis") on true positive rate values to infer the required sample size to reach a power of 0.8 with a fixed FDR for varying log-fold changes (Supplementary Fig. 4d–f). We estimated that a relative change of 1 (log2 scale) in abundant cell types (e.g., 1000 out of 5000 cells) can be determined with five samples, while the same relative change requires between 20 and 30 samples in a rare cell type (e.g., 125 out of 5000 cells) at an FDR level of 0.2. Notably, large relative changes (log-fold changes of 4) in rare cell types could be detected with less than ten samples. While this implies that for many situations only a few replicates are necessary, we would advise to increase the number of samples when detection of compositional changes in rare cell types is relevant.

scCODA identifies the FACS-verified decrease of B cells in supercentenarians. Next, we applied scCODA to a number of scRNA-seq data examples[1,3,4,6,19] (Fig. 3, Supplementary Figs. 5–9, and Supplementary Data 1). To confirm scCODA's applicability on real data with known ground truth, we first considered a recent study of age-related changes in peripheral blood mononuclear cells (PBMCs)[3], where cellular characteristics of supercentenarians ($n = 7$) were compared against the ones of younger controls ($n = 5$; Fig. 3a). The original study used a Wilcoxon rank-sum test and reported a significant decrease of B cells in supercentenarians, which is known from literature[20]. Moreover, the result was validated by FACS measurements. scCODA also identified B-cell populations as the sole affected cell type using CD16 + monocytes as a reference at an FDR level of 0.2. This suggests that scRNA-seq data indeed comprise enough information to study compositional changes, and that scCODA can correctly identify the experimentally validated age-related decrease of B cells even in low-sample regimes.

scCODA detects staining confirmed increase of disease-associated microglia in Alzheimer's disease on few replicates. Second, we analyzed the compositional changes of three microglia cell types in an Alzheimer's disease (AD) mouse model[19] (Fig. 3b and Supplementary Data 2). Here, the number of replicates of sorted cells from cortex and cerebellum was low ($n = 2$ per group), thus challenging standard statistical testing scenarios. In the cortex, scCODA identified statistically credible changes both in microglia 2 and disease-associated microglia (DAM) using the most abundant tissue-resident microglia 1 as reference cell type, or a credible change in microglia 1 when using one of the other two types of microglia as the reference. By contrast, scCODA detected no statistically credible change in the cerebellum, which is known to be unperturbed in AD. Keren-Shaul et al.[19] quantified the increase of DAM in the cortex of the AD mouse model via staining. While DAM localize in close proximity to amyloid-beta plaques and show a distinct inflammatory gene expression pattern, microglia 2 tend to represent an intermediate state between DAM and homeostatic microglia 1[19] (Supplementary Fig. 6). Therefore, our analysis with scCODA supports the contribution of DAM in AD. For comparison, ANCOM identified all three types of microglia as significantly changing in the cortex, and none in the cerebellum.

scCODA scales to large sample sizes and cell-type numbers. We next analyzed compositional changes of cell types in single-cell data from patients with ulcerative colitis (UC) compared to healthy donors[1]. Here, biopsy samples from the epithelium and the underlying lamina propria (Fig. 3c, Supplementary Data 3, and Supplementary Fig. 7) were enzymatically separated and subsequently analyzed with scRNA-seq, resulting in 51 cell types from 133 samples. The epithelium and the lamina propria represent two different compartments and were tested separately. However, some epithelial cells ended up in the lamina propria samples and vice versa. For testing, we summarized these cells as nonepithelial in the epithelium and as epithelial in the lamina propria (Fig. 3d, e). We then reanalyzed the data with the Dirichlet regression model used in Smillie et al.[1], leading to more statistically significant results compared to the original publication. Similar to the Dirichlet regression model, scCODA identified several statistically credible cell-type changes in healthy tissue compared to both non-inflamed and inflamed tissue in both epithelium and lamina propria at an FDR level of 0.2, using Immature Goblet cells and CD8 + intraepithelial lymphocytes (IELs) as automatically selected reference. Notably, we tested scCODA with different reference cell types (Supplementary Fig. 8), and did not detect credible changes for CD8 + IELs in the lamina propria with any reference cell type, backing up that CD8 + IELs are a good reference that does not change with respect to any other cell type. In the epithelium, both Dirichlet regression and scCODA identified significant and statistically credible changes, respectively, in the absorptive and secretory lineage, but scCODA also identified an increase in enteroendocrine cells. For comparison, ANCOM only identified significant changes in M cells (healthy vs inflamed) and enteroendocrine cells (healthy vs non-inflamed). M cells are lowly abundant and only 3 out of 16 inflamed samples had more than ten cells. When we compared the M-cell-positive subset of inflamed samples to healthy samples, though, M cells were indeed credibly increased. In the lamina propria, B-cell subpopulations showed several changes, e.g., a decrease of plasma B cells with disease (validated with stainings in Smillie et al.[1]), and an increase of follicular B cells. Moreover, consistent with our simulation studies demonstrating scCODA's higher sensitivity for lowly abundant cell types, scCODA uniquely detected statistically credible changes in several low-abundant immune cell populations. For instance, scCODA identified regulatory T cells (T_{reg}) to be more abundant in UC patients which is consistent with other studies[21]. Smillie et al. combined the results of their Dirichlet regression with two non-compositional tests, Fisher exact test and Wilcoxon rank-



sum test, to identify absolute changes in each population independently. Using such a two-stage procedure, Smillie et al. also reported changes in the low-abundant cell types such as T_{reg} cells. For comparison, ANCOM only identified significant changes in inflammatory fibroblasts (healthy vs inflamed), epithelial cells and pericytes (healthy vs non-inflamed), while all cell types in non-inflamed vs inflamed were reported as significantly changing. In contrast to Dirichlet Regression, scCODA reported credible changes in inflammatory fibroblasts (IAFs) for the healthy vs. inflamed case. Similar to M cells in the epithelium, IAFs form a responding subgroup within the inflamed donor group: While the cell type was almost absent in the control group, 13 out of 24 UC patient samples had more than five cells, indicating that ANCOM and scCODA are more likely to detect lowly abundant or absent cell types, where

Fig. 3 scCODA determines the compositional changes in a variety of examples. References are indicated in bold. **a** Boxplots of blood samples of supercentenarians ($n = 7$, dark blue) have significantly fewer B cells than younger individuals (control, $n = 5$, light blue), reference was set to CD16+ Monocytes, Hamiltonian Monte Carlo (HMC) chain length was set to 20,000 with a burn-in of 5000. Credible and significant results are depicted as colored bars (red: scCODA, brown: Wilcoxon rank-sum test (two-sided; Benjamini–Hochberg corrected)[3]). Results are in accordance with FACS data[3]. *P* values and effect sizes are shown in Supplementary Data 1**. b** Microglia associated with Alzheimer's disease (AD) are significantly more abundant in the cortex, but not in the cerebellum[19] ($n = 2$ in AD (dark blue) and wild-type (light blue) mice, respectively), HMC chain length was set to 20,000 with burn-in of 5000. *P* values and effect sizes are shown in Supplementary Data 2**. c**–**e** Changes in epithelium and lamina propria in the human colon[1] in ulcerative colitis (UC) ($n = 133$ from 18 UC patients, 12 healthy donors). Credible and significant results are depicted as colored bars (red: scCODA, green: two-sided *t* test of Dirichlet regression coefficients). Stars indicate the significance level (*adjusted $P < 0.05$, **adjusted $P < 0.01$, ****adjusted $P < 0.001$; Benjamini–Hochberg corrected). **c** Epithelium and Lamina propria are distinct tissues, which are studied separately. **d** Compositional changes from healthy (light blue) to non-inflamed (medium blue) and inflamed (dark blue) biopsies of the intestinal epithelium, HMC chain length was set to 150,000 with burn-in of 10,000. *P* values and effect sizes are shown in Supplementary Data 3**. e** Boxplots of compositional changes from healthy (light blue) to non-inflamed (medium blue) and inflamed (dark blue) biopsies in the lamina propria, HMC chain length was set to 400,000 with burn-in of 10,000. *P* values and effect sizes are shown in Supplementary Data 3**. f** Boxplots of compositional changes in bronchoalveolar cells in COVID-19 patients ($n = 4$ healthy (light blue), $n = 3$ mild (medium blue), $n = 6$ severe (dark blue) disease progression)[4]. Credible and significant results are depicted as colored bars (red: scCODA, orange: *t* test (two-sided; Benjamini–Hochberg corrected)), references for scCODA: Plasma (all pairwise comparisons between conditions), FDR at 0.2. Stars indicate the significance level (*: adjusted $P < 0.05$, **adjusted $P < 0.01$, ***adjusted $P < 0.001$; Benjamini–Hochberg corrected), HMC chain length was set to 80,000 with a burn-in of 10,000. *P* values and effect sizes are shown in Supplementary Data 4**. a**, **b**, **d**–**f** In all boxplots, the central line denotes the median, boxes represent the interquartile range (IQR), and whiskers show the distribution except for outliers. Outliers are all points outside 1.5 times of the IQR.

changes manifest in a subset of samples. On the other hand, only 10 out of 24 samples in the non-inflamed group had more than five IAFs, which was not enough for both methods to detect a credible change. We tested scCODA's performance to detect compositional changes when only a subset of samples exhibits a response (Supplementary Fig. 9). Consistent with the observations on IAFs, lowly abundant cell types show credible changes only when at least half of the samples change upon stimulation.

scCODA detects cell-type changes in COVID-19 patients that were not detected with non-compositional tests but confirmed in larger-scale studies. Next, we reanalyzed a recent COVID-19 single-cell study comparing compositional changes of major cell types in bronchoalveolar lavage fluid between healthy controls ($n = 4$), severe ($n = 6$) and moderate ($n = 3$) COVID-19 cases[4] using plasma as manually selected reference (Fig. 3f and Supplementary Data 4). The study originally reported significant differential changes in pDC's in healthy vs moderate and moderate vs severe, respectively, depletion in mDCs in severe vs healthy, and depletion of T cells in severe cases vs. moderate cases using a *t* test without multiple testing correction. Correcting for multiple testing resulted in only pDC's reported as significantly changing in healthy vs mild and mild vs severe cases, respectively. scCODA confirmed the differential change in T cells, and identified a credible increase in NK cells between mild vs healthy cases, credible depletions of T cells between moderate vs severe cases, as well as a credible increase of neutrophils in healthy and moderate vs severe at an FDR level of 0.2 using Plasma as reference. For comparison, ANCOM identified significant changes in mDCs between healthy and moderate, as well as neutrophils between healthy and moderate vs severe at alpha=0.2, respectively. The correlation of T-cell abundances with severity is well established and has been used as risk factors for severe cases[22,23]. A decrease of NK cells with COVID-19 severity was observed between recovered and diseased patients[23] in PBMC through FACS analysis. Finally, higher neutrophil proportions have been associated with severe outcomes[24] and are suspected to be the main drivers of the exacerbated host response[25], further confirming scCODA's findings.

scCODA accounts for the negative correlation structure for compositional changes and shows fewer false positives. Our final analysis considered a longitudinal scRNA-seq dataset from the small intestinal epithelium in mice, studying the effects of *Salmonella* and *Heligmosomoides polygyrus* infection on cell-type composition[6]. In contrast to the original Poisson regression data analysis[6], scCODA found only a single statistically credible increase in Enterocytes in *Salmonella* infected mice for an FDR level of 0.2 (Supplementary Fig. 10 and Supplementary Data 5). In addition, the Poisson model identified Tuft cells to be significantly affected after three and ten days of infection with *H. polygyrus*, while Enterocytes, Goblet, and early transit-amplifying cells were found to change significantly only after ten days of infection (Supplementary Fig. 10). All these changes could not be confirmed by scCODA at an FDR level of 0.2. For comparison, ANCOM did not find any significant changes for all three conditions, confirming its lack of power for datasets with few samples.

Discussion

In summary, using a comprehensive set of synthetic and scRNA-derived compositional datasets and application scenarios, we established scCODA's excellent performance for identifying statistically credible changes in cell-type compositions, while controlling for the false discovery rate. scCODA compared favorably to commonly used models for single-cell and microbiome compositional analysis, particularly when only a low number of experimental replicates are available. We believe this is due to the Bayesian nature of the model as it adequately accounts for the uncertainty of observed cell counts, automatically performs model selection, and does not rely on asymptotic assumptions. scCODA not only correctly reproduced previously discovered and partially FACS-verified compositional changes in recent scRNA-seq studies, but also identified additional cell-type shifts that were confirmed by independent studies, including T_{reg} cell enrichment in UC patients and neutrophils increase in severe COVID-19 cases. Using synthetic benchmarks, we confirmed that standard univariate tests, such as Poisson regression models, Beta-Binomial regression, or *t* tests are inadequate for cell-type analysis, since they do not account for the compositional nature of the data. While log-ratio transforms from compositional data analysis (such as the ALR used here) can partially mitigate these shortcomings, our Bayesian scCODA framework provided substantial performance improvements across all tested scenarios and is particularly preferable when only few replicates are available. Other methods from the field of microbiome data analysis,

such as ANCOM and ANCOM-BC, showed similar detection power, but could not adequately control the false discovery rate in the low-sample regimes.

While scCODA shows excellent performance in our simulation studies and applications, the current modeling framework possesses several limitations. In its present form, the scCODA framework requires pre-specified cell-type definitions which, in turn, hinge on statistically sound and biologically meaningful clustering assignments. In situations where crisp clustering boundaries are elusive, for instance, due to the presence of the transient developmental processes underlying the data, joint modeling of different resolution hierarchies[26] or modeling compositional processes[27,28] may help account for such continuities changes. Furthermore, scCODA assumes a log-linear relationship between covariates and cell abundance, which may be misspecified in some cases. Thus, scCODA may benefit from incorporating appropriate transformation models for the covariate data to achieve approximately log-linear relations. In its current form, scCODA does not model or infer any dependency structure among the cell compositions beyond the ones induced by the compositional effects. While more complex dependencies could, in principle, be included via additional hyperpriors, this would considerably increase the computational complexity and would require more efficient inference algorithms. Finally, scCODA does not model the response variability within a condition and thus cannot detect heterogeneities between samples in response to treatment or donor variability, as, e.g., in the data of UC patients[1]. This could be addressed by adding a novel covariate to inspect subsets of the data.

Overall, we believe that our scCODA framework offers an ideal starting point to model such advanced processes thanks to its hierarchical and extendable nature.

Methods

Model description. We seek to identify the credibly associated covariates X^{NxM} to observed cell counts Y^{NxK} of K cell types measured in a single-cell experiment with N samples and M covariates. We address this question with a Bayesian generalized linear multivariate regression framework using a Dirichlet-Multinomial model with a log-link function to account for the compositional nature and uncertainties in the observed data. Effects between covariates m and cell types k are hierarchically modeled using individual, normally distributed effects $\gamma_{m,k}$ with a covariate-specific scaling factor σ_m^2 [29,30]. For automatic model selection and identification of credibly associated covariates and affected cell types, we utilize a logit-normal prior as a continuous relaxation of the spike-and-slab prior[10] resulting in the following hierarchical model:

$$Y \sim \mathrm{DirMult}(\phi, \bar{y}) \tag{1}$$

$$\log(\phi) = \boldsymbol{\alpha} + X\beta \tag{2}$$

$$\alpha_k \sim \mathrm{N}(0,5) \quad \forall k \in [1,..,K] \tag{3}$$

$$\beta = \tau\tilde{\beta} \tag{4}$$

$$\tau_{m,k} = \frac{\exp(t_{m,k})}{1+\exp(t_{m,k})} \forall m \in [1,\dots,M],\ \forall k \in [1,\dots,K] \tag{5}$$

$$\frac{t_{m,k}}{50} \sim \mathrm{N}(0,1)\ \forall m \in [1,\dots,M],\ \forall k \in [1,\dots,K] \tag{6}$$

$$\tilde{\beta}_{m,k} = \sigma^2\gamma_{m,k} \forall m \in [1,..,M],\ \forall k \in [1,..,K] \tag{7}$$

$$\sigma_m^2 \sim \mathrm{HC}(1) \forall m \in [1,..,M] \tag{8}$$

$$\gamma_{m,k} \sim \mathrm{N}(0,1) \forall m \in [1,..,M], \forall k \in [1,..,K] \tag{9}$$

with N describing a Normal and HC a Half-Cauchy distribution following Polson et al.'s suggesting of hyperpriors for global scale parameters[31].

To prevent identifiability issues of the covariate parameters, we reparametrize the model and choose one cell type k as a reference, forcing its covariates $\beta_k = 0$ as in Maier et al.[11] (Methods—"Automatic reference selection").

Parameter inference is performed via Hamiltonian Monte Carlo (HMC) sampling using ten leapfrog steps per iteration with automatic step size adjustment according to Betancourt et al.[32]. Per default 20,000 iterations are performed with 5,000 iterations used as burn-in. The parameters $\alpha_k, \gamma_{m,k}$ are randomly initiated by drawing from standard normal priors. $t_{m,k}$ is always initialized with 0 to ensure unbiased model selection, while σ^2 is initialized with 1. If the data contains entries that are zero, a pseudocount of 0.5 is added to these zero counts to reduce numerical instabilities.

After parameter inference, we calculate the inclusion probability $P_{inc}(\beta_{k,m})$ of the covariates as follows:

$$P(\beta_{m,k}) = \frac{1}{H}\sum_{h=1}^{H} \mathbb{I}(|\beta_{m,k,h}| \geq 10^{-3}) \tag{10}$$

with H the number of HMC iterations and $\mathbb{I}$ the indicator function. To identify credibly associated covariates, we compare the calculated inclusion probabilities with a decision threshold c, which is determined *a posteriori* to control for the false discovery rate (Methods—"Spike-and-slab threshold determination"). For credible effects, we report the effect parameter $\beta_{m,k}$ as the mean over all MCMC samples where $\beta_{m,k}$ was nonzero.

Spike-and-slab threshold determination. To identify statistically credible effects, scCODA compares the posterior inclusion probability to a threshold c. As noted previously[33,34], Bayesian variable selection methods must control for multiplicity, to avoid an inflated number of false-positive associations. To this end, we use a direct posterior probability approach[9,35] to estimate the false discovery rate for a threshold value c.

By taking the posterior inclusion probability $P(\beta_{m,k})$ as an approximation for the certainty of a credible effect for each $\beta_{m,k}$, its complementary $1 - P(\beta_{m,k})$ approximates the probability of a type I error. For a threshold c, we now rank all $\beta_{m,k}$ by their type I error probability and obtain a set of credible effects $J(c) = \{\beta_{m,k} | 1 - P(\beta_{m,k}) \leq c\}$. Then, the approximate false discovery rate for the threshold is

$$\widehat{\mathrm{FDR}}(c) = \frac{\sum_{\beta_{m,k} \in J(c)} 1 - P(\beta_{m,k})}{|J(c)|}. \tag{11}$$

For a desired false discovery rate α, we now set the optimal threshold c' to include as many effects as possible, without the approximate FDR exceeding α:

$$c' = \min_{0<c<1; \widehat{\mathrm{FDR}}(c)<\alpha} c \tag{12}$$

Finally, $J(c')$ is the set of credible effects that is reported by scCODA.

Automatic reference selection. The compositional nature of scRNA-seq population data only allows statements about changes in abundance with respect to a reference group[5,8,11]. One way of defining such a reference is by selecting one cell type and interpreting changes to the other cell types with respect to this reference type. scCODA achieves this by forcing all effects on the reference cell type to be zero. The reference should therefore be set to a cell type that is known to be unaffected by the covariates.

However, such a cell type might not be known a priori. To alleviate this problem, scCODA offers an automatic reference selection that aims at selecting a cell type that is mostly unchanged in relative abundance, implying that the abundance of the reference cell type is stable over all samples. This is achieved by selecting the cell type that has the least dispersion of relative abundance over all samples, while being present in at least a fraction t of the samples:

$$K_{ref} = \mathrm{argmin}_{k \in \{1\dots K\}} \mathrm{Disp}(Y'_{.,k}) \text{ s.t. } \frac{|\{n: Y_{n,k}>0\}|}{N} \geq t. \tag{13}$$

Here, Y' is the relative abundance of cell counts. The additional condition on the reference cell type occurring in almost every sample is necessary to prevent very rare cell types from being selected, where small random changes in cell counts have a large impact on the relative abundance. Therefore, we recommend setting $t = 0.95$, meaning that the reference cell type has to be present in at least 95% of samples. If no such cell type exists, this constraint can be relaxed by lowering t.

We now show how the choice of the reference cell type can influence the results of scCODA. As an example, we use the ulcerative colitis Lamina propria data from Smillie et al.[1], comparing healthy and non-inflamed samples. We applied scCODA to this data 37 times, setting each cell type as the reference once (FDR level 0.05). Supplementary Fig. 8 shows the credible effects and effect size for each reference. For reference cell types that were mostly unchanged, i.e., were almost never found to be differentially abundant in the other runs, the found credible effects are largely consistent. On the other hand, cell types that were assigned a large negative effect (CD4+ activated Fos-lo, plasma cells) found significantly less credible effects when used as the reference, as the null level for the change is already negative. Taking epithelial cells, the only increasing cell type, as the reference led to the largest number of credible negative effects in other cell types. This shows that the reference cell type can have a large impact on the results of scCODA and should therefore be chosen with care.

Credible intervals. To measure the certainty of scCODA's credible effects, we calculate high-density intervals[36] for each effect parameter $\beta_{m,k}$. Due to the spike-and-slab prior formulation, posterior samples of β are naturally zero-inflated, with the extent depending on each effect's inclusion probability.

To counteract this bias, we, therefore, report credible intervals under the assumption that the effect in question is included in the model by calculating the high-density interval for each effect only across MCMC samples where the corresponding spike-and-slab variable was not 0:

$$\widehat{HDI}(\beta_{m,k}) = HDI(\beta_{m,k}|\tau_{m,k} > 0). \quad (14)$$

Supplementary Fig. 11 shows how excluding the non-credible samples changes the 95% HDI for the example of healthy vs. non-inflamed samples of ulcerative colitis from the Lamina Propria[1]. While excluding the zero samples from the HDI calculation influences the HDI of most cell types only marginally, some high-density intervals become slightly wider (CD69- mast cells) or shift away from zero (cycling B cells). The average width of 95% HDIs increases only slightly from 0.92 to 0.97, though. Note that generally Bayesian high-density intervals are relatively large due to the MCMC sampling uncertainty.

Simulation description. We carried out all benchmark studies by repeatedly generating compositional datasets $y \in N^{(n_0+n_1)xK}$ that have similar properties as the data from scRNA-seq experiments. For all synthetic datasets, we assumed a case-control setup with n_0 and n_1 samples in the two groups and K cell types, as well as a constant number of cells $\bar{y}$ in each sample.

We generated the synthetic datasets rowwise, with each row a sample of a Multinomial (MN) distribution $\boldsymbol{y_i} = \mathrm{MN}(\boldsymbol{\alpha}, \bar{y})$, and the probability vector $\boldsymbol{\alpha}$ a softmax transformation of a multivariate normal (MVN) sample: $\boldsymbol{\alpha} = \mathrm{softmax}(\mathrm{MVN}(\boldsymbol{\mu}, \Sigma))$. We always used a covariance matrix of $\Sigma = 0.05\, Id_K$, which mimics the variances observed in the experimental data of Haber et al., while assuming no correlation between the cell types besides the compositional effects[6].

In the power, heterogeneous response, and runtime analysis benchmarks, the mean vector μ for each sample was calculated from the mean abundance of the first cell type in control samples (no effect) μ_0, and the mean change in abundance of the first cell type between the two groups μ'. All other cell types were modeled to be equally abundant, leading to $\boldsymbol{\mu} = \log(\mu_0, \frac{\bar{y}-\mu_0}{K-1}, \frac{\bar{y}-\mu_0}{K-1}, \dots)$ for control samples, and $\boldsymbol{\mu} = \log(\mu_0 + \mu', \frac{\bar{y}-(\mu_0+\mu')}{K-1}, \frac{\bar{y}-(\mu_0+\mu')}{K-1}, \dots)$ for samples in the other group.

For the model comparison benchmark, we also included effects on two different cell types. For this, we assumed $\boldsymbol{\mu} = \log(1000, 1000, \dots, 1000)$ for all control samples, and an increase of $\boldsymbol{\mu'} = (\mu'_1, \mu'_2)$ on the first two cell types, leading to $\boldsymbol{\mu} = \log(1000 + \mu'_1, 1000 + \mu'_2, \frac{K \cdot 1000 - (2000+\mu'_1+\mu'_2)}{K-2})$.

For all benchmark studies, we defined sets of values for all parameters mentioned above and generated r datasets for every possible parameter combination. We then applied scCODA with the last cell type chosen as reference to each synthetic dataset. For the model comparison benchmark (Methods—"Model comparison"), we analyzed the results at FDR levels of 0.05 and 0.2. The overall benchmark (Methods—"Power analysis"), heterogeneous response benchmark (Methods—"Analysis of heterogeneous response groups") and runtime analysis (Methods—"Runtime analysis") were carried out with an expected FDR level of 0.05.

The sets of generation parameters were as follows:

- Model comparison (Fig. 2, Methods—"Model comparison"):

$$K = \{5, 10, 15\};$$

$$n_0 = n_1 = \{1, 2, 3, 4, 5\}(\text{only balanced setups} - n_0 = n_1);$$

$$\bar{y} = K \cdot 1000;$$

$$\mu_0 = 1000;$$

$$\boldsymbol{\mu'} = (0, 500); (0, 1000); (0, 2000); (500, 1000); (500, 2000); (1000, 2000);$$

$$r = 20;$$

- Power analysis (Supplementary Fig. 4, Methods—"Power analysis"):

$$K = 5;$$

$$n_0 = n_1 = \{1, 2, ..., 10\}(\text{also imbalanced setups});$$

$$\bar{y} = 5000;$$

$$\mu_0 = \{20, 30, 50, 75, 115, 180, 280, 430, 667, 1000\};$$

$$\boldsymbol{\mu'} = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 400, 600, 800, 1000\};$$

$$r = 10;$$

- Heterogeneous response groups (Supplementary Fig. 9, Methods—"Analysis of heterogeneous response groups"):

$$K = 5;$$

$$n_0 = n_1 = 20;$$

$$\bar{y} = 5000;$$

$$\mu_0 = \{1, 100, 1000\};$$

$$\boldsymbol{\mu'} = \{500, 1000, 2000\};$$

$$r = 20;$$

- Runtime analysis (Supplementary Fig. 12, Methods—"Runtime analysis"):

$$K = \{5, 10, 15, \dots 50\};$$

$$n_0 = n_1 = \{5, 10, 15, 20\}(\text{only balanced setups} - n_0 = n_1);$$

$$\bar{y} = 100,000;$$

$$\mu_0 = \frac{1}{K};$$

$$\boldsymbol{\mu'} = \{1\};$$

$$r = 20;$$

Power analysis. To be able to estimate the required sample sizes for an intended MCC, we fitted a quasibinomial regression model with log-linker function using log sample size, log absolute change in cell count, log-fold change, and all pairwise interactions using the simulation results at different fixed FDR levels (FDR=[0.05, 0.1, 0.2]; Methods—"Simulation description").

We performed a backward model selection with repeated tenfold cross-validation to reduce the feature set. The final model consisted of log sample size, log-fold change, log-fold change, and the interaction effects between log total sample size and log absolute cell count change as well as log absolute cell count and log-fold change, which was expected given that we observed an interaction effect between these two variables in the raw benchmark results (Supplementary Fig. 4d–f).

With the fitted model, we inverse estimated the required log total sample size x_{ss} with fixed TPR y_{tpr}, log-fold change x_{fc}, and log absolute cell count change x_{cc} as:

$$x_{ss} = \frac{-(\alpha + \beta_{fc}x_{fc} + \beta_{fc,cc}x_{fc}x_{cc} - y_{tpr})}{(\beta_{ss} + \beta_{ss,cc}x_{cc})} \quad (15)$$

With this formula, we estimated the sample size for a fixed power of 0.8 across changing log-fold changes between [0.01, 5] and the fraction of cell-type sizes to total cell counts between [0.01, 0.2] for the same fixed FDR levels.

Analysis of publicly available datasets

Single-cell RNA-seq data of PBMCs from supercentenarians. We downloaded the processed single-cell RNA-seq count matrices comprising PBMCs of seven supercentenarians and five younger controls from http://gerg.gsc.riken.jp/SC2018/. Read counts were log-transformed and PCA embedded using the first 50 PCs. Leiden clustering was used to cluster cells into major groups. Following the described analysis in Hashimoto et al.[3], we annotated the major cell types including T cells characterized by *CD3* and T-cell receptor (*TRAC*) expression, B cells characterized by *MS4A1* (*CD20*) and *CD19* expression, natural killer cells characterized by *KLRF1* expression, monocytes characterized by *CD14* and *FCGR3A* (*CD16*) expression, respectively, and erythrocytes characterized by *HBA1* expression, and determined their cell counts per sample (Supplementary Fig. 5). All analysis steps were carried out using Scanpy v.1.5.1.

Single-cell RNA-seq data of microglia in Alzheimer's disease (AD) mouse model. We downloaded the raw single-cell RNA-seq count matrices (deposited at GEO, accession code GSE98969) comprising immune cells isolated from the mouse brain in wild-type (WT) and AD mice[19]. The complete dataset with all samples consists of 37,248 cells. We filtered out ERCC spike-ins before computing the quality metrics of all cells. We then excluded 12,053 cells with less than 500 UMI counts and 11,065 genes, which were not expressed. We subsequently normalized by library size with target sum 10,000 counts (CPM normalization) and log+1 scaled. Following the analysis of Keren-Shaul et al.[19], we selected the samples of six-month-old mice from AD and WT, which have not been sorted by brain region, resulting in 9,196 cells. It must be noted that Keren-Shaul et al. reported 8,016 cells when they first annotated immune cells in 6-month-old mice (see Fig. 1 in Keren-Shaul et al.). We evaluated batch effects based on the clustering results and visual inspection of the UMAP plots, where none of the samples clustered separately in any of the clusters, which is, in this case,

sufficient to obtain cell types. We clustered the data using Louvain clustering with resolution 1 and annotated cell types using the previously reported marker genes as microglia 1 (*CTSD, CD9, HEXB, CST3*), microglia 2–3 (*LPL, CST*), granulocytes (*CAMP, S100a9*), T/NK cells (*S100a4, NKG7, Trbc2*), B cells (*RAG1, CD79b, CD74*), monocytes (*S100a4, CD74*), perivascular macrophages (*CD74, CD163, MRC1*) (see Supplementary Fig. 6). We subsequently sub-clustered the microglia population into three clusters, assigning the labels microglia 1, 2, and 3, respectively. Similar to Keren-Shaul et al., we assigned the region-sorted samples of AD and WT mouse model ($n = 2$ per region) with a k-nearest neighbor classifier ($k = 30$). We then evaluated the number of unassigned cells, performed another round of Louvain clustering, and assigned the remaining cells based on the majority vote for the clustering result, i.e., when unassigned cells clustered predominantly with microglia 1, they were all assigned to microglia 1. The obtained proportions of microglia subpopulations are in accordance with the previously reported proportions. All analysis steps were carried out using Scanpy v.1.5.1.

Single-cell RNA-seq data of ulcerative colitis in human donors. We used the annotated single-cell RNA-seq data of the colon epithelium from 12 healthy donors and 18 patients with chronic inflammation[1]. From healthy donors, samples from two adjacent locations were taken. From patients, biopsies from inflamed and adjacent normal tissue ("non-inflamed") were taken. Further, the biopsies were separated by enzymatic digestion into the epithelium ("Epi") and the lamina propria ("LP") before single-cell RNA-sequencing. The study comprises a total of 365,492 transcriptomes from 133 samples. The data were downloaded from Single Cell Portal (accession ID SCP259). The analysis code and description were provided at https://github.com/cssmillie/ulcerative_colitis.

The original study annotated all cell types together, resulting in 51 different cell types. However, some cell types that are originally located in the LP have been found in the epithelial samples and vice versa. For the differential composition analysis of the Epi and LP, we considered the nonepithelial and epithelial cell types, respectively, as one group. Therefore, we tested the changes in 16 cell types in the Epi and 37 cell types in LP. In addition, we reanalyzed the data using the Dirichlet regression model as in Smillie et al.[1] (with R package DirichletReg v.0.7-0 in R v.3.5.2). Importantly, we realized that Smillie et al. summed up the counts of the same replicates (as described in the analysis scripts in https://github.com/cssmillie/ulcerative_colitis), while we consider every replicate as an independent. Overall, we have data from 29 donors (61 samples, where 24 healthy, 21 non-inflamed, 16 inflamed) in Epi and data from 30 donors (72 samples, where 24 each healthy, non-inflamed, and inflamed, respectively) in LP. Specifically, we tested chain lengths of 20,000, 40,000, 80,000, and 150,000 iterations with a burn-in of 10,000 in the Epi case, while we tested chain lengths of 200,000, 400,000, and 800,000 iterations with a burn-in of 10,000 in the LP case due to the larger number of cell types in LP compared to Epi.

Single-cell RNA-seq data of bronchoalveolar immune cells in patients with COVID-19. We used the annotated single-cell RNA-seq data of the bronchoalveolar lavage fluid cells from three patients with moderate COVID-19 progression, six patients with severe COVID-19 progression, four healthy donors, and a publicly available sample[4]. The cell-type annotations of all samples were provided at https://github.com/zhangzlab/covid_balf.

Single-cell RNA-seq data of small intestinal epithelial cells infected with different bacteria. Annotated single-cell transcriptomics data of epithelial cells from the small intestine of mice infected with three different bacterial conditions were downloaded from Single Cell Portal (accession ID SCP44). The data consisted of a control group of four mice (3,240 cells total) and three groups of two mice each, measured after 2 days for *Salmonella* (1,770 cells total), as well as three (2,121 cells total) and ten days (2,711 cells total) after *H. polygyrus* infection, respectively.

Model comparison. We compared scCODA's ability to correctly identify significant compositional changes in a setting typical for single-cell experiments to other methods recently used in scRNA-seq analysis and approaches from the field of microbial population analysis. We applied all methods to each of the 5,000 datasets generated for the comparison analysis (Methods—"Simulation description") and recorded which of the cell types each method found to be differentially abundant between the two groups. We then compared these results to the ground truth assumption from the data-generation process via binary classification metrics (credible vs. non-credible changes). We chose Matthews' correlation coefficient as our primary metric, as it best accounts for the numerical imbalance between the two groups. Details on the individual differential abundance testing methods can be found in Supplementary Table 2. We also investigated the False discovery rate and sensitivity (true positive rate) for each method for a more detailed performance analysis.

Furthermore, we performed sensitivity analysis via the receiver-operating characteristic and precision-recall curve. The different methods use different metrics (e.g., *P* values) that can be thresholded to obtain the sensitivity curves. The thresholding metric, AUC score, and average precision score for each method are listed in Supplementary Table 1.

Analysis of heterogeneous response groups. In certain cases, only a fraction of the samples in a treatment group show a response to the stimulus. To quantify the sensitivity of scCODA in such scenarios, we conducted another benchmark study. We simulated datasets as before, assuming that either a rare or an abundant cell type was increasing by a significant margin in the treatment group (Methods—"Simulation description"). To mimic a partial response to the covariate, we defined treatment groups where the affected cell type was increased in (5%, 10%, … 100%) of the samples, while the rest of the samples followed the distribution of the control group.

Independent of the abundance of a cell type, scCODA detected the effects only if a relatively large share of the samples was responsive to the condition. For abundant cell types (base count $\mu_0 = 100$ or 1000), a response share of about 40% was enough to achieve reliable detection, while for very rare cell types (base count $\mu_0 = 1$), more than half of the samples needed to show a response. If the share of responding samples was 70% or higher, scCODA reliably detected the effects (Supplementary Fig. 9).

We therefore conclude that scCODA is robust to small amounts of non-responding samples within a condition. However, scCODA does not detect compositional changes that only manifest in a minority share within a condition. In that case, the changes will be considered as outliers rather than credible effects.

Runtime analysis. To benchmark the execution time and scalability of scCODA with the size of the data, we generated a collection of 800 datasets with an increasing number of cell types and samples (Methods—"Simulation description"). The generation parameters were chosen such that the typical dimensions of scRNA-seq datasets are covered by the benchmark.

scCODA uses HMC sampling for parameter inference. Therefore, the most important factor in runtime is the duration of one HMC sampling step. To isolate the HMC sampling process from the model initialization and post-sampling analysis steps, we applied scCODA twice to each dataset, sampling chains of length 1,000 and 2,000, respectively. We measured the execution time for both instances and divided the time difference by 1,000—the difference in chain length—to gain an estimate for the execution time per sampling iteration (Supplementary Fig. 12). All operations were executed on an Intel(R) Xeon(R) Gold 6126 processor. The memory consumption of a single run of scCODA in default settings should not exceed 2 GB.

For five cell types, datasets of all tested sample sizes require about 0.0025 s per HMC iteration on average. The time per iteration increased linearly with the number of cell types for all sample sizes. This effect is more pronounced for larger sample sizes, with 40 total samples (20 per group) and 50 cell types requiring the longest average time per step of about 0.0035 s, while the average runtime per step for datasets with five samples was always below 0.0027 s. Thus, running scCODA with the default number of 20,000 HMC iterations on any dataset of typical size should produce results within a few minutes.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The synthetic benchmark datasets and results have been deposited on Zenodo at https://doi.org/10.5281/zenodo.4305907. The single-cell datasets can be found in their respective public repositories. The supercentenarians PBMC dataset by Hashimoto et al. can be found at http://gerg.gsc.riken.jp/SC2018, while the Alzheimer's mouse microglia dataset by Keren-Shaul et al. can be accessed at GEO under GSE98969. The single-cell ulcerative colitis dataset by Smillie et al. can be downloaded from the Single-Cell Portal (Accession ID SCP259) and its accompanying analysis code and description from https://github.com/cssmillie/ulcerative_colitis. The processed single-cell data of bronchoalveolar immune cells in patients with COVID-19 by Liao et al. is publicly available at https://github.com/zhangzlab/covid_balf. The single-cell data of small intestinal epithelial cells infected with different bacteria is available from Single Cell Portal (accession ID SCP44).

Code availability

The method has been implemented in Python 3.8 using Tensorflow = 2.3.2[37], Tensorflow-Probability = 0.11[38], ArviZ >= 0.10[39], numpy >= 1.19, and Scanpy >= 1.5[12]. The Power Analysis was performed using caret package[40] (R 4.1). Source code has been deposited on Github at https://github.com/theislab/sccoda[41]. All code to reproduce the presented analyses can be found on Github at https://github.com/theislab/scCODA_reproducibility[42]. All tested methods have been integrated into a unifying Python API that can directly interact with Scanpy and Anndata.

Received: 11 January 2021; Accepted: 1 November 2021;
Published online: 25 November 2021

References

1. Smillie, C. S. et al. Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* **178**, 714–730.e22 (2019).
2. Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).

3. Hashimoto, K. et al. Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians. *Proc. Natl Acad. Sci. USA* **116**, 24242–24251 (2019).
4. Liao, M. et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26**, 842–844 (2020).
5. Aitchison, J. The statistical analysis of compositional data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **44**, 139–160 (1982).
6. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
7. Cao, Y. et al. scDC: single cell differential composition analysis. *BMC Bioinforma.* **20**, 721 (2019).
8. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* **8**, 2224 (2017).
9. Wadsworth, W. D. et al. An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinforma.* **18**, 94 (2017).
10. Thomson, W., Jabbari, S., Taylor, A. E., Arlt, W. & Smith, D. J. Simultaneous parameter estimation and variable selection via the logit-normal continuous analogue of the spike-and-slab prior. *J. R. Soc. Interface* **16**, 20180572 (2019).
11. Maier, M. J. DirichletReg: Dirichlet regression for compositional data in R. *Research Report Series, Vienna University of Economics and Business.* **125**, 1-26 (2014).
12. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
13. Lin, H. & Peddada, S. D. Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* **11**, 3514 (2020).
14. Mandal, S. et al. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26**, 27663 (2015).
15. Gloor, G. ALDEx2: ANOVA-like differential expression tool for compositional data. *ALDEX Man. Modular* **20**, 1–11 (2015).
16. Martin, B. D., Witten, D. & Willis, A. D. Modeling microbial abundances and dysbiosis with beta-binomial regression. *Ann. Appl. Stat.* **14**, 94–115 (2020)
17. Mallick, H. et al. Multivariable association discovery in population-scale meta-omics studies. *PLoS Comput. Biol.* **17**, e1009442 (2021)
18. Hawinkel, S., Mattiello, F., Bijnens, L. & Thas, O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.* **20**, 210–221 (2019).
19. Keren-Shaul, H. et al. A unique microglia type associated with restricting development of Alzheimer's disease. *Cell* **169**, 1276–1290.e17 (2017).
20. Franceschi, C., Monti, D., Sansoni, P. & Cossarizza, A. The immunology of exceptional individuals: the lesson of centenarians. *Immunol. Today* **16**, 12–16 (1995).
21. Holmén, N. et al. Functional CD4+CD25high regulatory T cells are enriched in the colonic mucosa of patients with active ulcerative colitis and increase with disease activity. *Inflamm. Bowel Dis.* **12**, 447–456 (2006).
22. Du, R.-H. et al. Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: a prospective cohort study. *Eur. Respir. J.* **55**, 2050524 (2020).
23. Zhang, B. et al. The dynamics of immune response in COVID-19 patients with different illness severity. *J. Med. Virol.* **93**, 1070–1077 (2020).
24. Wang, D. et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *J. Am. Med. Assoc.* **323**, 1061–1069 (2020).
25. Barnes, B. J. et al. Targeting potential drivers of COVID-19: neutrophil extracellular traps. *J. Exp. Med.* **217**, e20200652 (2020).
26. Bien, J., Yan, X., Simpson, L. & Müller, C. L. Tree-aggregated predictive modeling of microbiome data. *Sci. Rep.* **11**, 14505 (2021).
27. Pawlowsky-Glahn, V., Egozcue, J. J. & Tolosana-Delgado, R. *Modeling and Analysis of Compositional Data* (John Wiley & Sons, 2015).
28. Äijö, T., Müller, C. L. & Bonneau, R. Temporal probabilistic modeling of bacterial compositions derived from 16 S rRNA sequencing. *Bioinformatics* **34**, 372–380 (2018).
29. Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. Non-centered Parameterisations for Hierarchical Models and Data Augmentation. *Bayesian Statistics* **7**, pp. 307–326 (Oxford University Press, 2003).
30. Papaspiliopoulos, O., Roberts, G. O. & Sköld, M. A general framework for the parametrization of hierarchical models. *Stat. Sci.* **22**, 59–73 (2007).
31. Polson, N. G. & Scott, J. G. On the half-cauchy prior for a global scale parameter. *Bayesian Anal.* **7**, 887–902 (2012).
32. Betancourt, M. J., Byrne, S. & Girolami, M. Optimizing the integrator step size for Hamiltonian Monte Carlo. Preprint at https://arxiv.org/abs/1411.6669 (2014).
33. Scott, J. G. & Berger, J. O. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annal Stat.* **38**, 2587–2619 (2010).
34. Muller, P., Parmigiani, G. & Rice, K. FDR and Bayesian Multiple Comparisons Rules. *Johns Hopkins University, Dept. of Biostatistics Working Papers*. **115**, 1–15 (2006)
35. Newton, M. A., Noueiry, A., Sarkar, D. & Ahlquist, P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155–176 (2004).
36. Kruschke, J. K. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan* (Academic Press, 2015).
37. Abadi, M. et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. Preprint at https://arxiv.org/abs/1603.04467 (2016)
38. Dillon, J. V. et al. TensorFlow distributions. Preprint at https://arxiv.org/abs/1711.10604 (2017).
39. Kumar, R., Carroll, C., Hartikainen, A. & Martin, O. ArviZ a unified library for exploratory analysis of Bayesian models in Python. *J. Open Source Softw.* **4**, 1143 (2019).
40. Kuhn, M. et al. caret: Classification and Regression Training. Astrophysics Source Code Library ascl:1505.003 (2015)
41. Büttner, M., Ostner, J., Müller, C. L., Theis, F. J. & Schubert, B. scCODA is a Bayesian model for compositional single-cell data analysis. https://github.com/theislab/scCODA, https://doi.org/10.5281/zenodo.5578025 (2021).
42. Büttner, M., Ostner, J., Müller, C. L., Theis, F. J. & Schubert, B. Reproducibility repository—scCODA is a Bayesian model for compositional single-cell data analysis. https://github.com/theislab/scCODA_reproducibility, https://doi.org/10.5281/zenodo.5578002 (2021).

Acknowledgements

We would like to thank Dr. Malte Luecken for his support in the initial design of the study, as well as Karin Hrovatin and Lisa Sikkema for their support in developing synthetic data-generation methods and testing the implementation of scCODA. We also thank Dr. Fabian Scheipl for his suggestions for defining credible intervals on spike-and-slab models. F.J.T. acknowledges financial support by the Bavarian Ministry of Science and the Arts in the framework of the Bavarian Research Association "ForInter" (Interaction of human brain cells). B.S acknowledges financial support by the Postdoctoral Fellowship Program of the Helmholtz Zentrum München.

Author contributions

M.B. and B.S. conceived of the study. J.O. developed scCODA and conducted the benchmarking study. M.B., J.O., and B.S. analyzed the data. C.L.M. helped design the model comparison. B.S. and F.J.T. supervised the study and model development. B.S. designed the benchmarking study and conducted the power analysis. M.B., J.O., C.L.M., and B.S. wrote the manuscript. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

F.J.T. reports receiving consulting fees from Roche Diagnostics GmbH and Cellarity Inc., and an ownership interest in Cellarity, Inc. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-27150-6.

Correspondence and requests for materials should be addressed to C. L. Müller or B. Schubert.

Peer review information *Nature Communications* thanks Ilya Korsunsky and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

scCODA is a Bayesian model for compositional single-cell data analysis

Büttner M.[1+], Ostner J.[1,2+], Müller CL.[1,2,3]*, Theis FJ.[1,4,5†], Schubert B.[1,4†*]

[1] Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

[2] Department of Statistics, Ludwig-Maximilians-Universität München, München, Germany

[3] Center for Computational Mathematics, Flatiron Institute, New York, New York, USA

[4] Department of Mathematics, Technische Universität München, Garching bei München, Germany

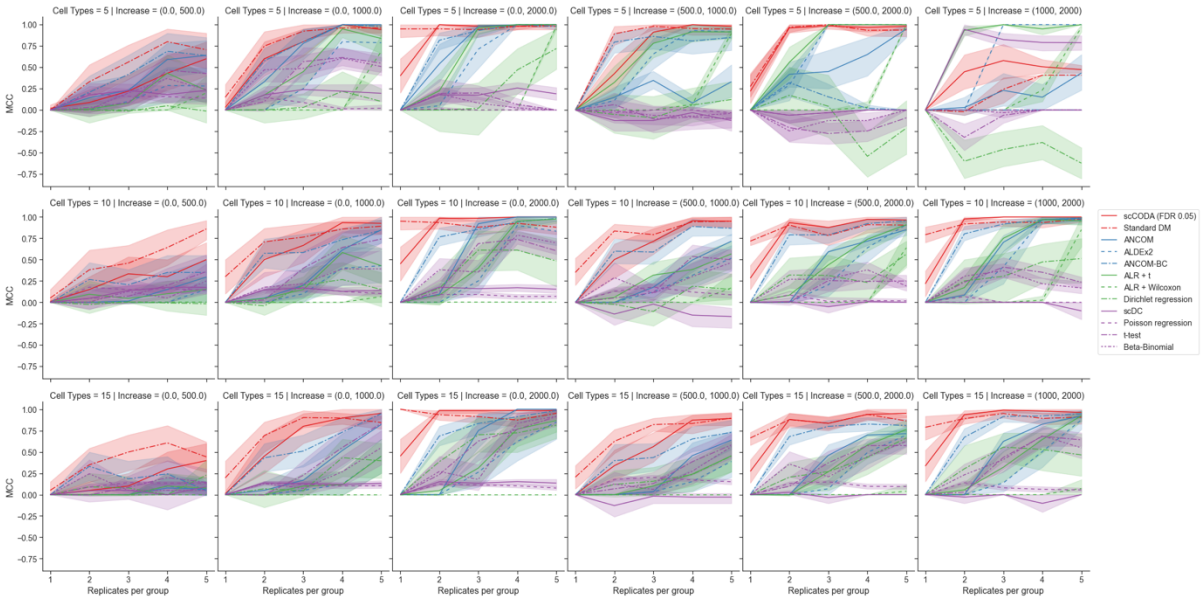
[5] TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany

[+]These authors contributed equally.

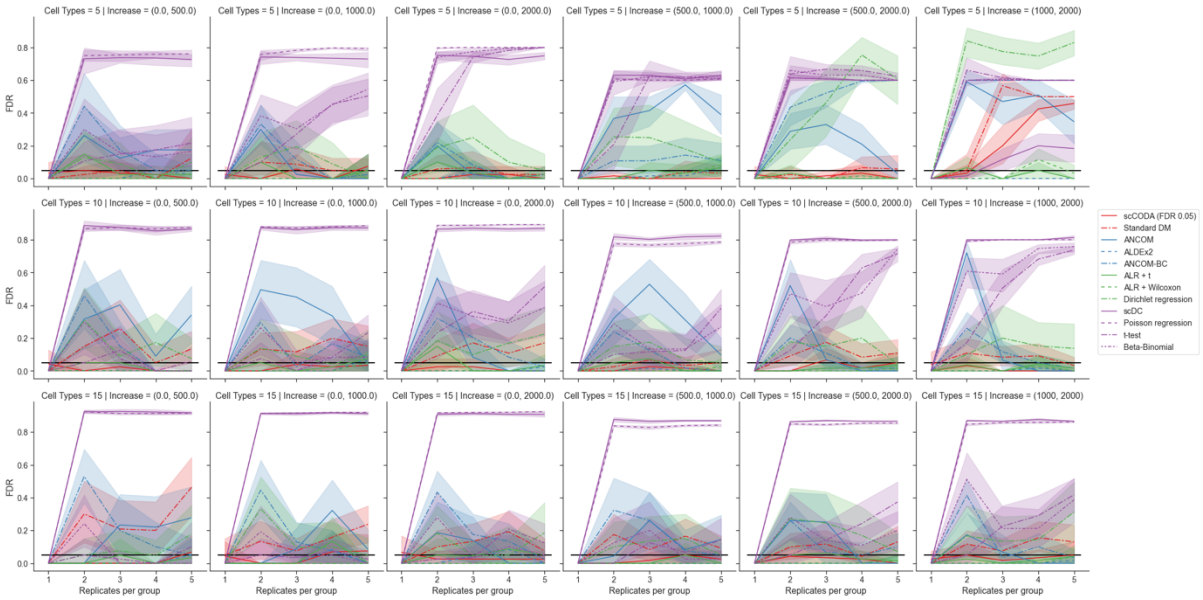
[†]These authors jointly supervised this work.

*Correspondence: christian.mueller@stat.uni-muenchen.de; benjamin.schubert@helmholtz-muenchen.de

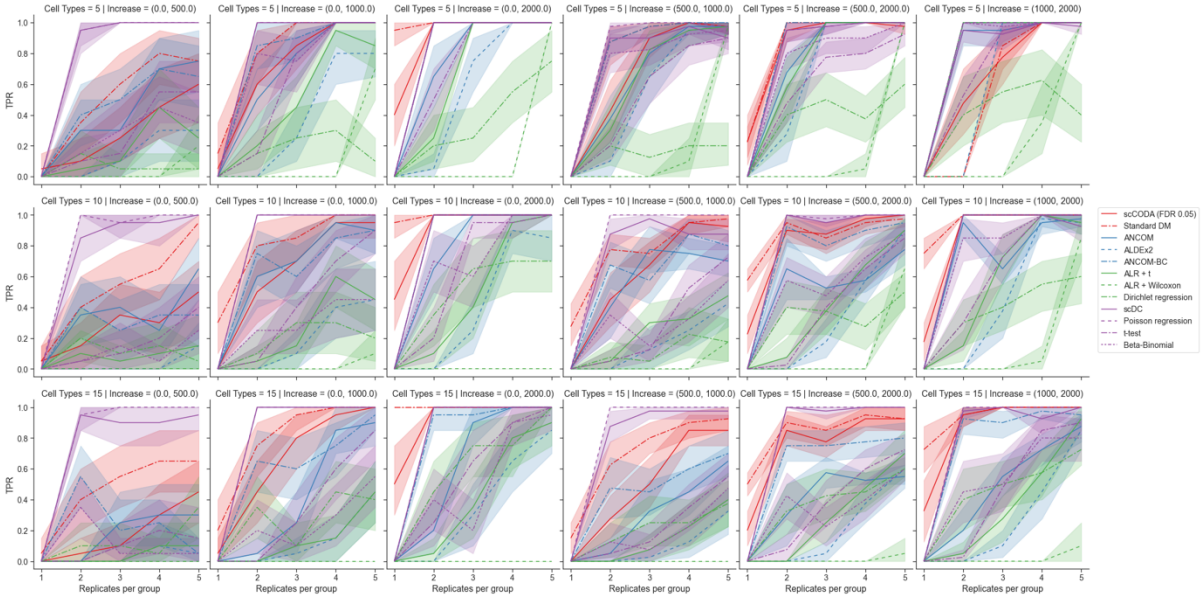
Supplementary Figures



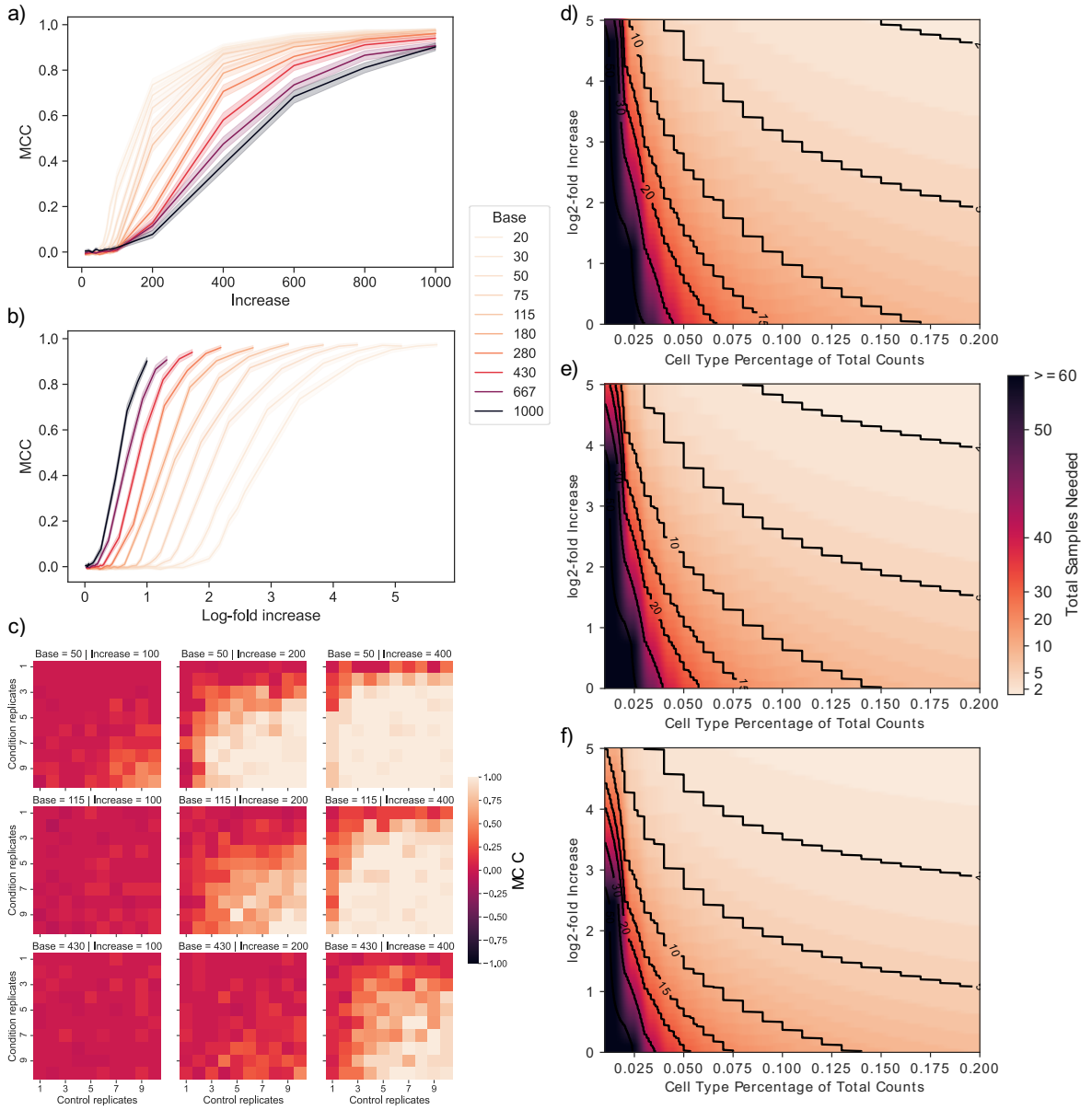
Supplementary Figure 1: Comparison of model performance by sample size measured by MCC, separated by number of cell types and effect size. The “Increase” value denotes the expected absolute change in the first two cell types between control and case samples. Error bands denote the 95% confidence interval around the mean.



Supplementary Figure 2: Comparison of model precision by sample size measured by FDR, separated by number of cell types and effect size. The "Increase" value denotes the expected absolute change in the first two cell types between control and case samples. Error bands denote the 95% confidence interval around the mean. The nominal FDR level of 0.05 for all methods is indicated with a horizontal black line. The case of 5 cell types and an increase of (1000, 2000) shows much higher FDR values, because the reference cell type has an expected count of 0 in the case samples.

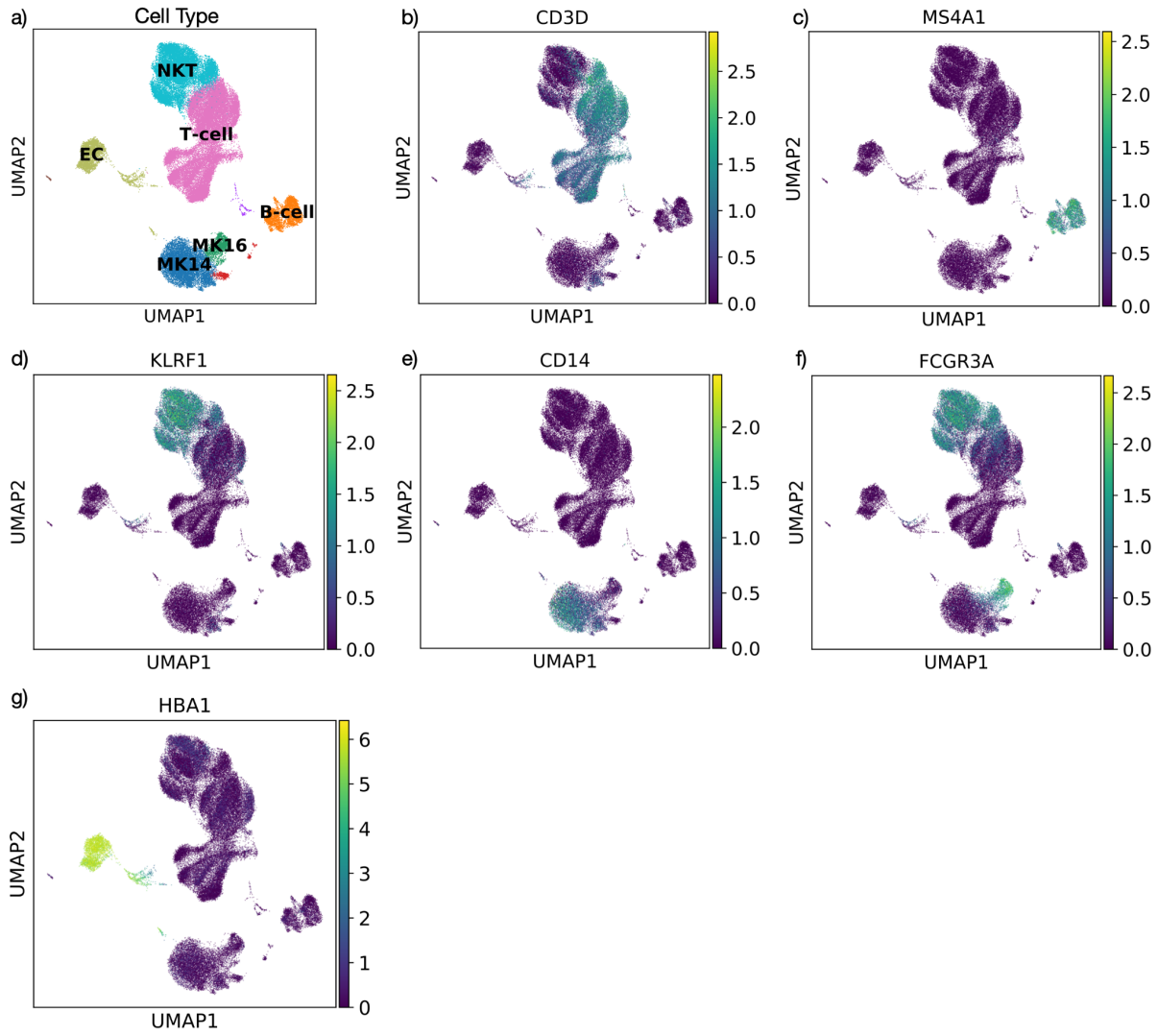


Supplementary Figure 3: Comparison of model sensitivity by sample size measured by TPR, separated by number of cell types and effect size. The “Increase” value denotes the expected absolute change in the first two cell types between control and case samples. Error bands denote the 95% confidence interval around the mean.

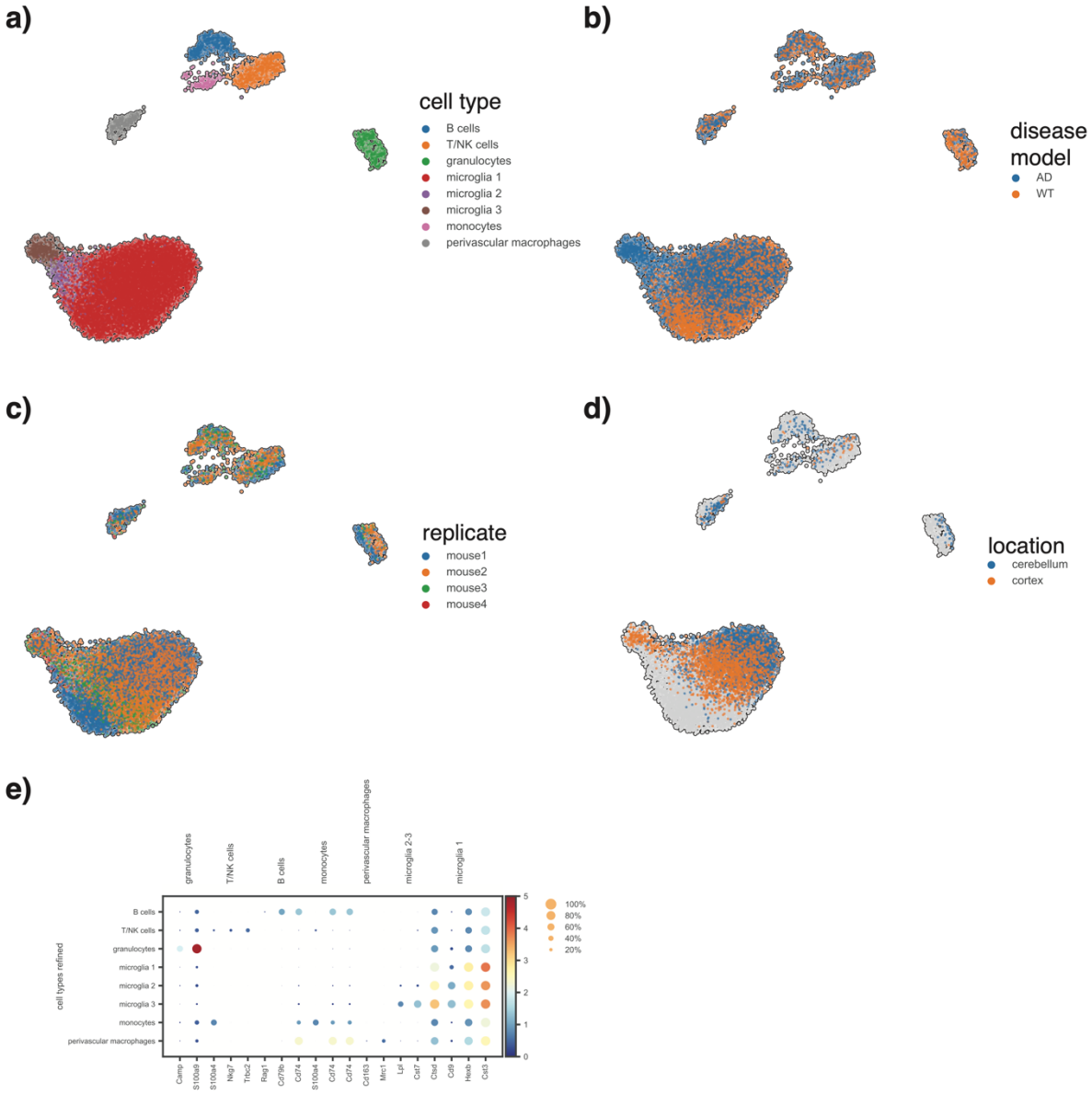


Supplementary Figure 4: Benchmark evaluation results for overall benchmark **(Methods - Simulation description)**. **(a)** The performance of scCODA (measured by MCC) depends on the amount of change in abundance. The "Base" value represents the mean cell count of the only differentially abundant cell type in the control group samples. For cell types with higher initial abundance, the absolute (count) change must be higher to reliably detect changes in abundance. Error bands denote the 95% confidence interval around the mean. **(b)** For cell types with higher initial abundance, scCODA can detect smaller relative (log2-fold) changes between the two groups. Error bands denote the 95% confidence interval around the mean. **(c)** Example

performance of scCODA depending on sample size of both groups. Less abundant cell types need a smaller absolute increase to be reliably detected as differentially abundant. The shaded areas in **(b-c)** represent 95% confidence intervals **(d-f)** Total samples per group needed to achieve an expected sensitivity of 0.8, depending on base value and increase of the differentially abundant cell type **(Methods - Power analysis)** for fixed FDR levels of **(d)** FDR=0.05, **(e)** FDR=0.1, and **(f)** FDR=0.2.

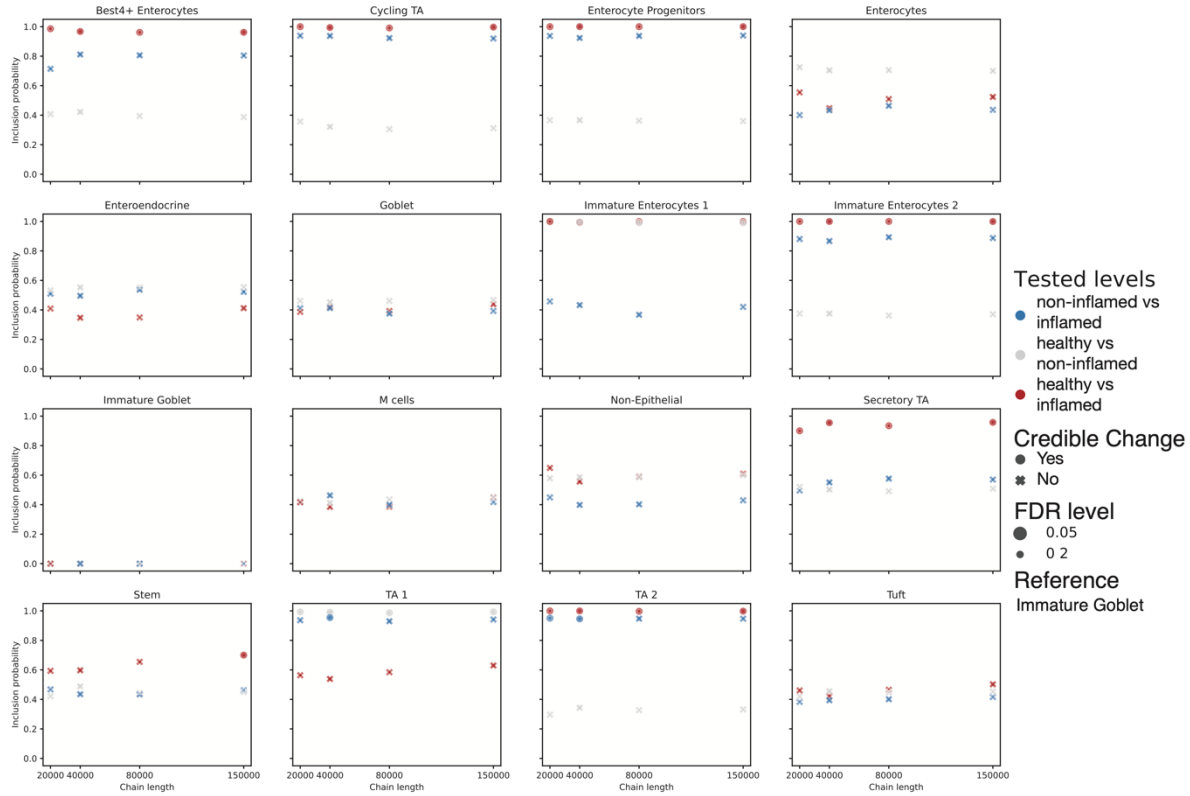


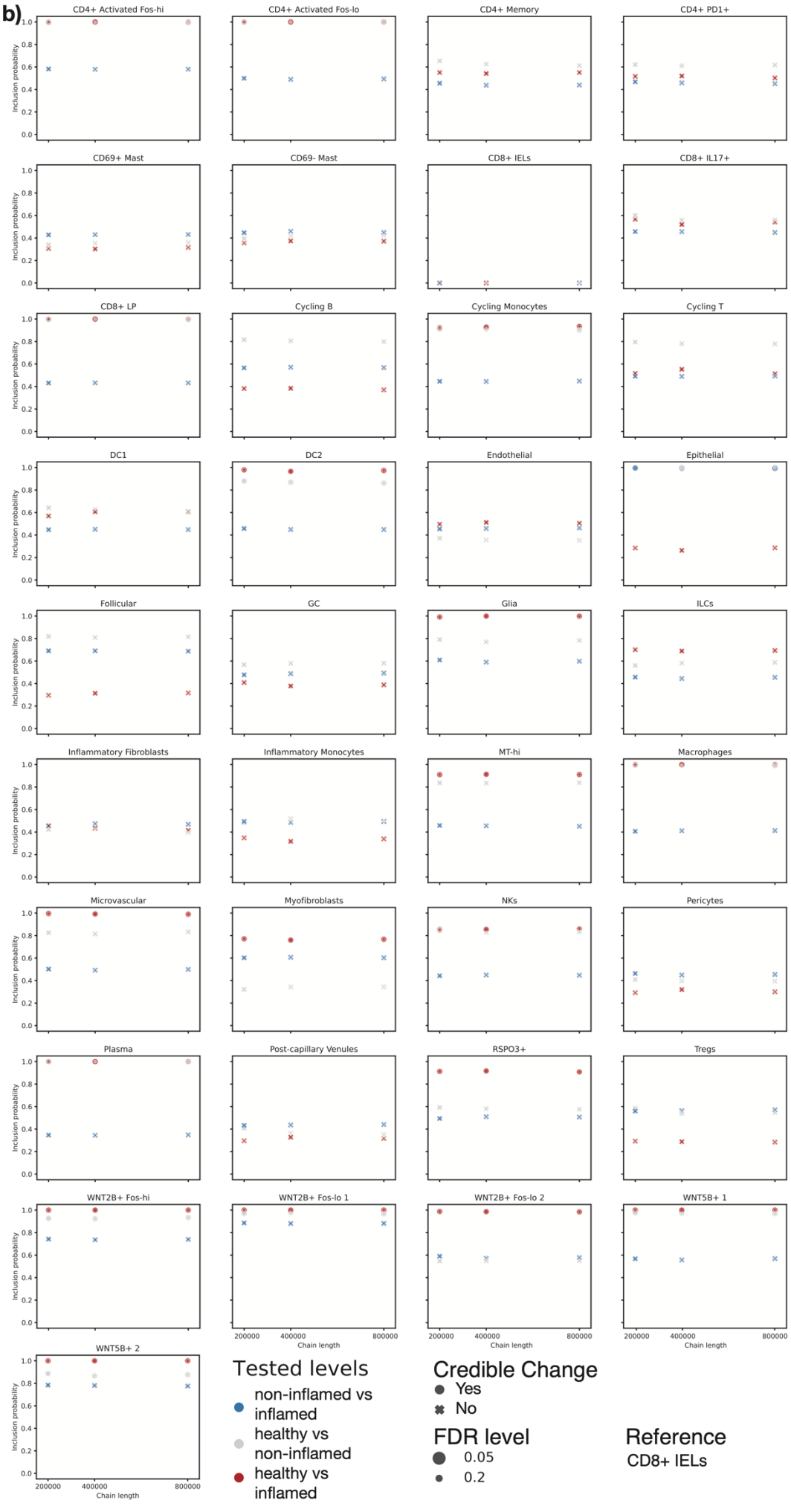
Supplementary Figure 5: Re-analysis of supercentenarian data of Hashimoto et al.[3]. **(a)** Final annotation of major cell types. **(b-g)** expression pattern of *CDR3* identifying T-cells, *MS4A1* identifying B-cells, KLRF*1* natural killer cells (NKT), *CD14* and *FCGR3A* (CD16) Monocyte subtypes (CD14+, CD16+, denoted as MK14 and MK16), and *HBA1* Erythrocytes (EC).

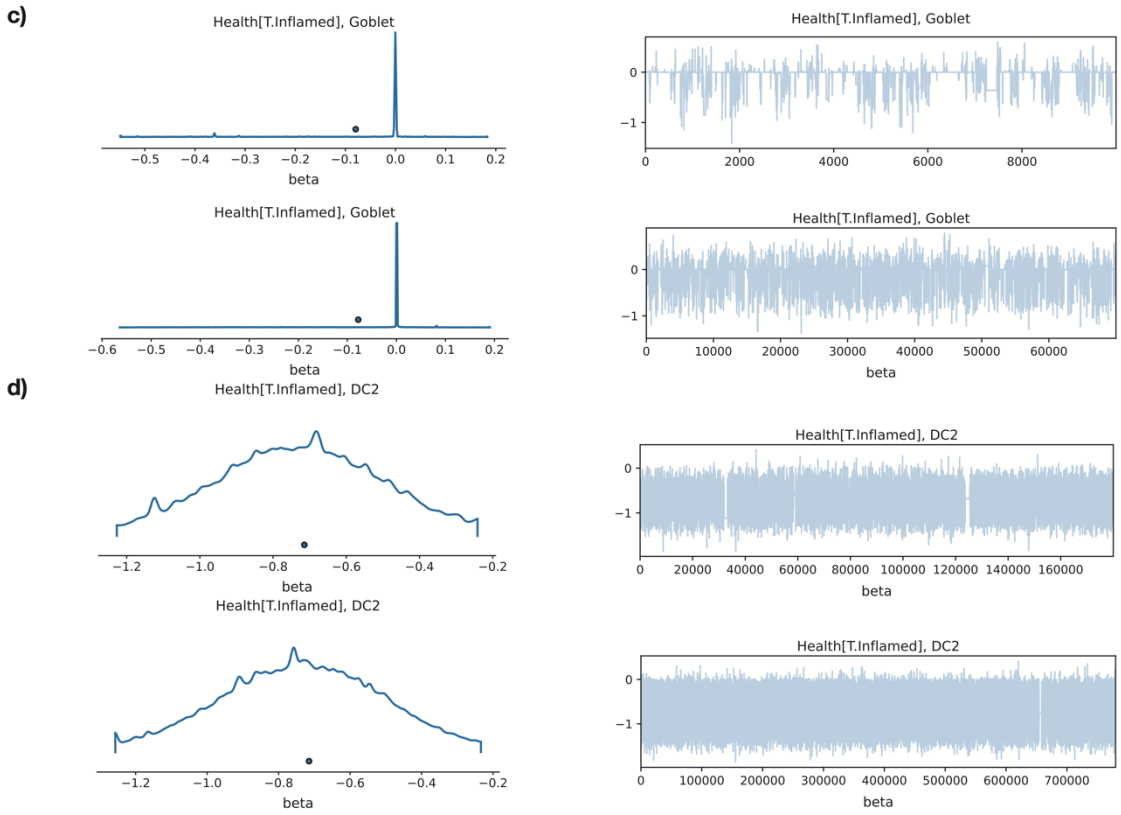


Supplementary Figure 6: Re-analysis of microglia data in Alzheimer's disease (AD) mouse model[19]. **(a)** Joint cell type annotation of cells. **(b)** Cell distribution in the both wild type (WT) and AD mouse models. **(c)** Distribution of cells from different replicates does not indicate strong batch effects. **(d)** Location of cells sorted from cortex and cerebellum. Location of grey cells was not reported. **(e)** Dot plot of marker gene expression of the annotated cell populations **(a)**.

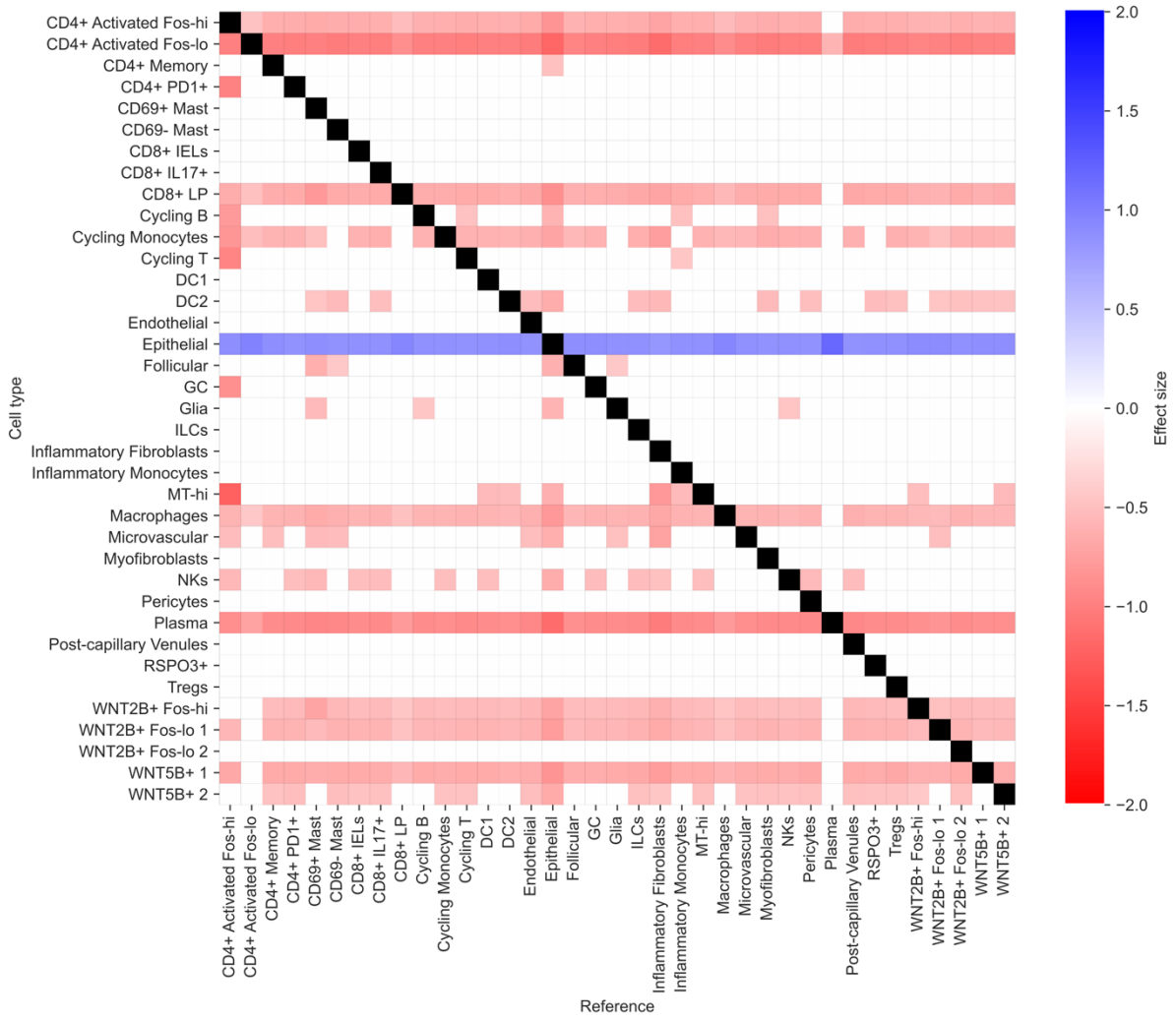
a)


Best4+ Enterocytes
Cycling TA
Enterocyte Progenitors
Enterocytes
Enteroendocrine
Goblet
Immature Enterocytes 1
Immature Enterocytes 2
Immature Goblet
M cells
Non-Epithelial
Secretory TA
Stem
TA 1
TA 2
Tuft
Inclusion probability
0.0
0.2
0.4
0.6
0.8
1.0
20000 40000
80000
150000
Chain length
Tested levels
non-inflamed vs inflamed
healthy vs non-inflamed
healthy vs inflamed
Credible Change
Yes
No
FDR level
0.05
0 2
Reference
Immature Goblet

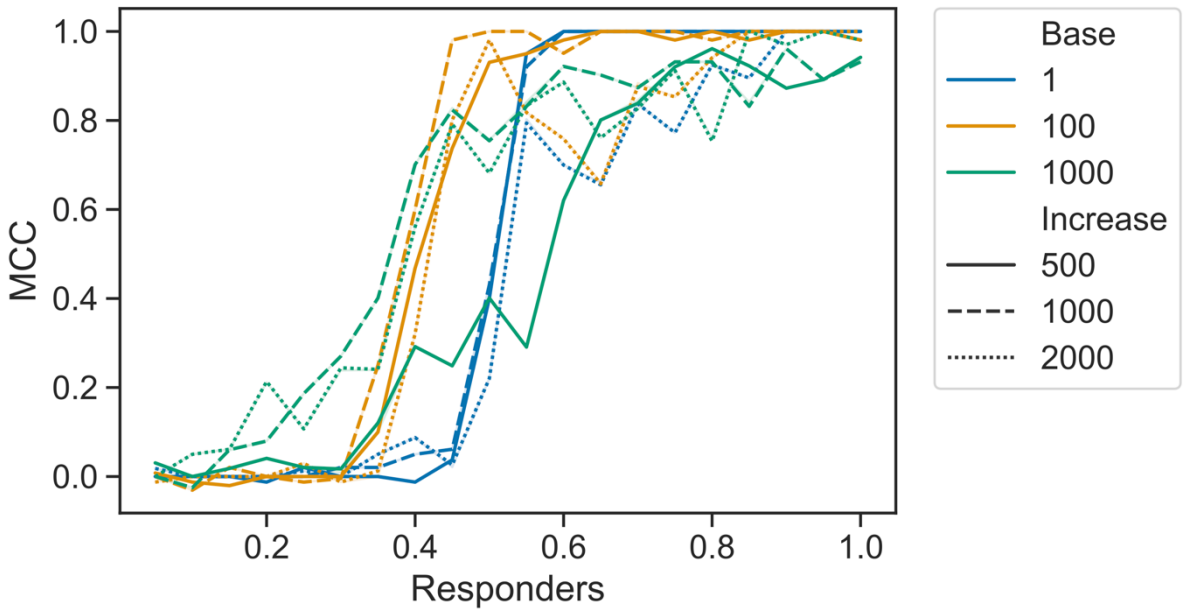

b)
CD4+ Activated Fos-hi
CD4+ Activated Fos-lo
CD4+ Memory
CD4+ PD1+
CD69+ Mast
CD69- Mast
CD8+ IELs
CD8+ IL17+
CD8+ LP
Cycling B
Cycling Monocytes
Cycling T
DC1
DC2
Endothelial
Epithelial
Follicular
GC
Glia
ILCs
Inflammatory Fibroblasts
Inflammatory Monocytes
MT-hi
Macrophages
Microvascular
Myofibroblasts
NKs
Pericytes
Plasma
Post-capillary Venules
RSPO3+
Tregs
WNT2B+ Fos-hi
WNT2B+ Fos-lo 1
WNT2B+ Fos-lo 2
WNT5B+ 1
WNT5B+ 2
Inclusion probability
0.0
0.2
0.4
0.6
0.8
1.0
200000
400000
800000
Chain length
Tested levels
non-inflamed vs inflamed
healthy vs non-inflamed
healthy vs inflamed
Credible Change
Yes
No
FDR level
0.05
0.2
Reference
CD8+ IELs



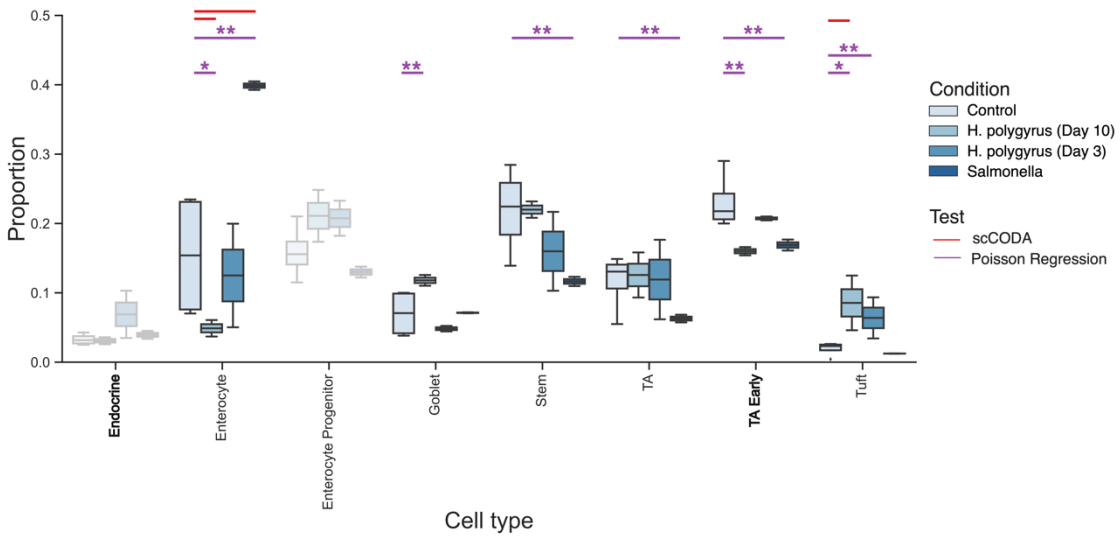
Supplementary Figure 7: Convergence of HMC sampling for many cell types (in data of Smillie et al.). **(a-b)** Inclusion probabilities for pairwise tests in the epithelium **(a)** and lamina propria **(b)** of healthy donors and patients of UC. Colors depict the tested levels; symbols depict the credibility of the changes. The effect of the reference is set to zero. **(c-d)** Density plots (left panels) and trace plots (right panels) of different chain lengths for the parameter inference in Goblet cells comparing healthy and inflamed samples (reference Immature Goblet) **(c)** and DC2 comparing healthy and inflamed samples (reference CD8+ IELs) **(d)**.



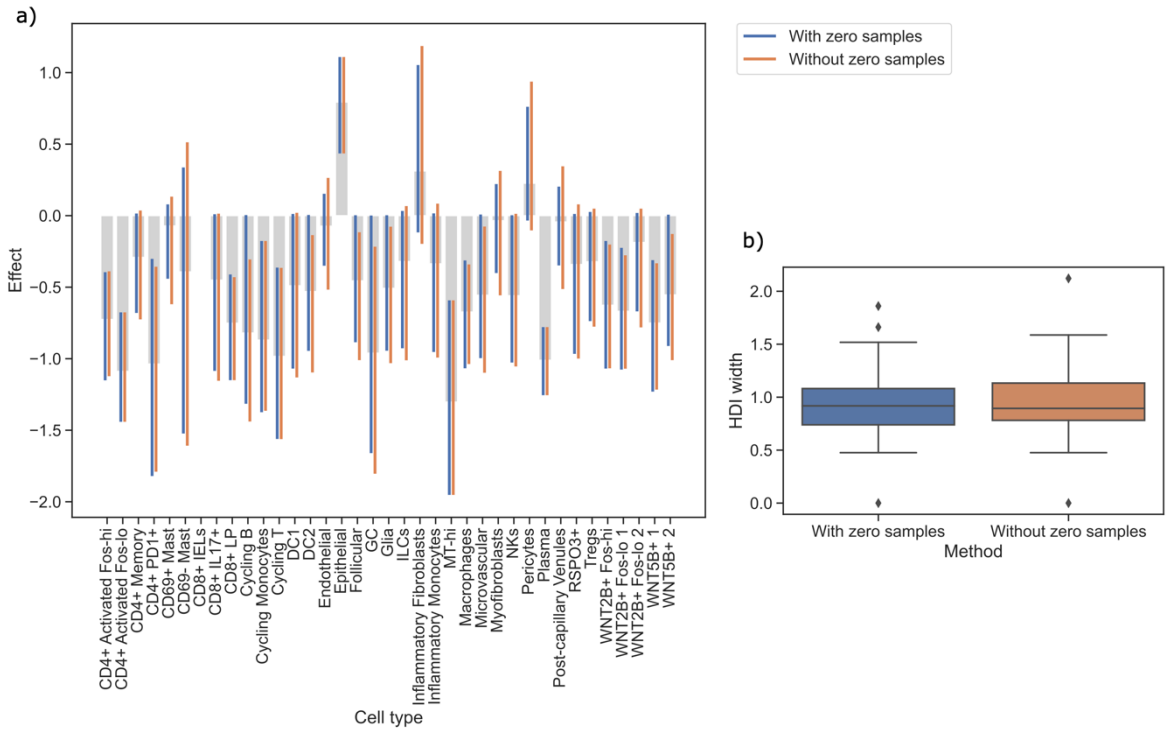
Supplementary Figure 8: Credible effects depend on the reference cell type in scCODA. Blue and red areas depict positive and negative credible effects, white areas show no credible effect. The reference cell type itself is colored black. Using a cell type as the reference that is often characterized with a large decrease (CD4+ Activated Fos-lo, Plasma cells) leads to less credible decreases being found. Using Epithelial cells, the only credibly increasing cell type, as the reference, leads to a larger number of negative effects. Data: Ulcerative colitis data from the Lamina propria (Healthy vs. non-inflamed)[1]



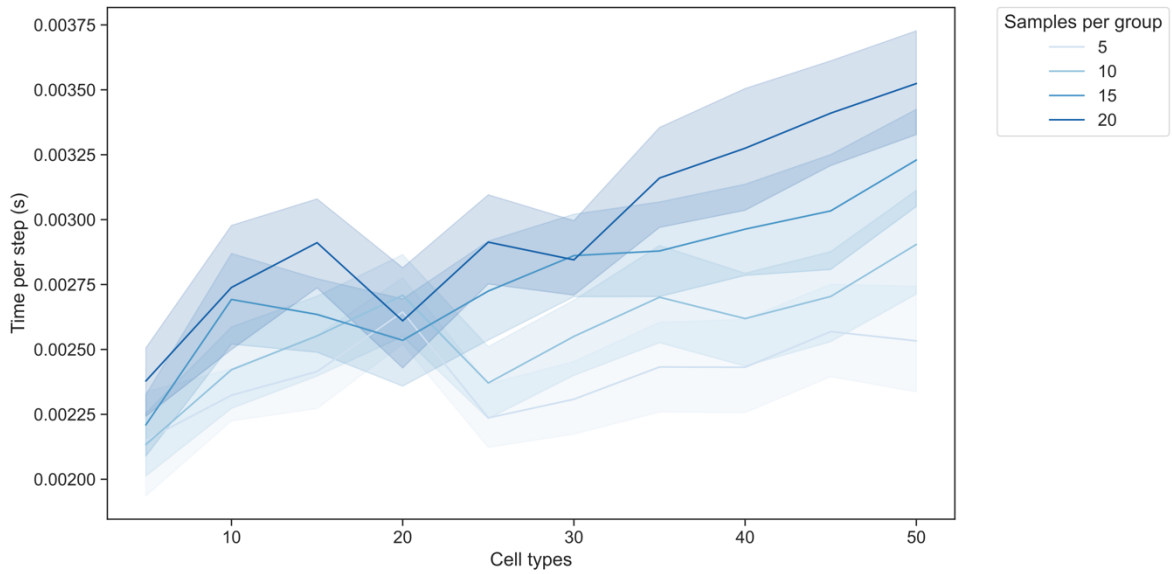
Supplementary Figure 9: Benchmarking results on response heterogeneity in a condition **(Methods - Analysis of heterogeneous response groups)**. The "Base" value indicates the mean count of the affected cell type in the control group, the "Increase" value represents the absolute increase between conditions. The x-axis shows the fraction of treatment samples that were simulated to respond to the condition. Only if more than half of the samples responded to the treatment, scCODA was able to reliably detect the effect, even in very rare cell types.



Supplementary Figure 10: Compositional analysis of Haber et al.[6] on the response to pathogen infection in the small intestinal epithelium of the mouse. Significant and credible results in comparison to the control population (n=4 animals) are depicted as colored bars (Red: scCODA, purple: Dirichlet regression), stars depict the significance of the Poisson regression model carried out by Haber et al.[6] (*: adjusted $p<10^{-5}$, **: adjusted $p<10^{-10}$). The reference cell type for scCODA was determined automatically to be Endocrine cells (*Salmonella* (n=2 animals) and *H. polygyrus (Day10)* (n=2 animals)), and early transit-amplifying (TA Early) cells (*H.Polygyrus (Day 3)* (n=2 animals)), respectively. In all box plots, the central line denotes the median, boxes represent the interquartile range (IQR), and whiskers show values within. P-values and effect sizes are shown in **Supplementary Data 5.** In all box plots, the central line denotes the median, boxes represent the interquartile range (IQR), and whiskers show the distribution except for outliers. Outliers are all points outside 1.5 times of the IQR.



Supplementary Figure 11: High (95%) posterior density intervals (HDI) for effect parameters on ulcerative colitis data from the Lamina propria (Healthy (n=24 samples) vs. non-inflamed (n=24 samples)). We compare two ways of calculating the HDI - with (blue) and without (orange) including the MCMC samples where the corresponding spike-and-slab variable is zero. **(a)** Interval range (blue and orange bars) of both methods for each cell type. The grey bars show the posterior mean (including zero samples) for each effect. CD8+ IELs have no effect, since they were used as the reference. **(b)** Boxplots of HDI width (difference between upper and lower interval boundary) across all cell types. In the box plots, the central line denotes the median, boxes represent the interquartile range (IQR), and whiskers show the distribution except for outliers. Outliers are all points outside 1.5 times of the IQR.



Supplementary Figure 12: Runtime analysis benchmark **(Methods - Runtime analysis)**. The time per HMC step (in seconds) is dependent on the number of cell types and the number of samples in the two treatment groups. The shaded areas depict the 95% confidence intervals around the mean. Generally, each HMC iteration takes longer for larger datasets. This effect is approximately linear in the number of cell types, with a less steep increase in runtime for datasets with fewer samples.

Supplementary Tables

Supplementary Table 1: Sensitivity analysis of differential abundance testing methods. AUC score from Receiver operating characteristic (**Fig. 2a**), Average precision score from precision-recall curve (**Fig. 2b**). All analyses were performed according to **Supplementary Table 2**.

Method	**ROC thresholding parameter**	**AUC**	**Average precision score**
scCODA	Inclusion probability	0.99	0.94
scDC	p-value	0.56	0.2
ANCOM	W-statistic	0.77	0.65
ALDEx2	p-value	0.9	0.77
ANCOM-BC	p-value	0.94	0.70
ALR+t-test	p-value	0.95	0.85
ALR+Wilcoxon	p-value	0.93	0.72
Dirichlet regression	p-value	0.7	0.31
Poisson regression	p-value	0.44	0.16
t-test	p-value	0.84	0.44
Beta-Binomial	p-value	0.84	0.41

Supplementary Table 2: Methods and configurations used in the benchmark comparison. Wrappers around implementations of all methods for easy use are implemented in the scCODA package. The Package column denotes the implementation that is called in scCODA.

Method	**Implementation details**	**Parameters**	**Package**
scCODA	Our proposed method	Reference cell type always set to the last component; FDR level 5%	scCODA package, version 0.1.3
Standard Dirichlet-Multinomial	Fully Bayesian model: Log-linear model on components of a Dirichlet-Multinomial distribution. Selection of a reference cell type. HMC inference setup identical to scCODA. Effects are credible if 0 is not included in the high-density interval	Reference cell type always set to the last component; High density interval: 95%	scCODA package, version 0.1.3
scDC	Single-cell differential composition analysis[7] Number of bootstrap samples generated for each data set: 100; no subject effects in linear model *Note: This method did not give results for all datasets. The erroneous results were left out of the analysis*	False discovery rate: 5%	R-package scdney[7], version 0.1.5
ANCOM	Analysis of composition of microbiomes[14]; Used test: t-test; Holm-Bonferroni multiplicity correction (all recommended settings)	False discovery rate: 5%	Python-package scikit-bio[42], version 0.5.6

ALDEx2	ANOVA-Like Differential Expression tool for high throughput sequencing data[43]. Reference cell type set to the last component instead of the geometric mean; testing via t-test (Benjamini-Hochberg-corrected)	False discovery rate: 5%	R-package ALDEx2[15], version 1.22
ANCOM-BC	Analysis of compositions of microbiomes with Bias correction[13]; Holm correction of p-values (recommended)	False discovery rate: 5%	R-package ANCOMBC[13], version 1.0.5
ALR+t-test	Additive log-ratio transform of data; t-test (two-sided) on all components; Benjamini-Hochberg correction of p-values	Reference component: Last cell type; False discovery rate: 5%	Python-package scipy[44], version 1.6.1
ALR+Wilcoxon	Additive log-ratio transform of data; Wilcoxon-rank-sum test (two-sided) on all components; Benjamini-Hochberg correction of p-values	Reference component: Last cell type; False discovery rate: 5%	Python-package scipy[44], version 1.6.1
Dirichlet regression	Default settings: One-sample t-test of Dirichlet regression coefficients	Significance level 5%	R-Package DirichletReg[11], version 0.7
Poisson regression	Poisson regression model used by Haber et al. [6]; Benjamini-Hochberg correction of p-values	False discovery rate: (as used by Haber et al. [6])	Python-package statsmodels[45], version 0.12.1
t-test	t-test (two-sided) on all components of untransformed data; Benjamini-Hochberg correction of p-values	False discovery rate: 5%	scipy[44], version 1.6.1
Beta-Binomial	Variance estimation only for more than 2 samples per group possible; Test statistic: Likelihood-ratio (recommended for small sample sizes); Benjamini-Hochberg correction of p-values	False discovery rate: 5%	R-package corncob[16], version 0.2.0

A.2. tascCODA: Bayesian Tree-Aggregated Analysis of Compositional Amplicon and Single-Cell Data

Contributing article

Ostner, J., Carcy, S., and Müller, C. L. (2021). tascCODA: Bayesian Tree-Aggregated Analysis of Compositional Amplicon and Single-Cell Data. *Front. Genet. 12, 766405.* doi: `https://doi.org/10.3389/fgene.2021.766405`

Replication code

Source code for this contribution has been deposited on Github at `https://github.com/theislab/sccoda`. The scripts used for data analysis and benchmark data generation can be found at `https://github.com/bio-datascience/tascCODA_reproducibility`. Supplemental data can be downloaded from zenodo (`https://zenodo.org/records/5302136`).

Copyright information

Author contributions

J.O. developed tascCODA and conducted the simulation studies and real data analysis. S.C. processed the 16S rRNA sequencing data and provided biological context. C.L.M. supervised the work. J.O. and C.L.M. conceived the statistical model, designed the simulation and out-of-sample prediction studies and wrote the manuscript. All authors read and approved the final manuscript.

frontiers in Genetics

METHODS
published: 07 December 2021
doi: 10.3389/fgene.2021.766405

tascCODA: Bayesian Tree-Aggregated Analysis of Compositional Amplicon and Single-Cell Data

Johannes Ostner[1,2], *Salomé Carcy*[2,3†] *and Christian L. Müller*[1,2,4*]

[1]Department of Statistics, Ludwig-Maximilians-Universität München, Munich, Germany, [2]Institute of Computational Biology, Helmholtz Zentrum München, Munich, Germany, [3]Department of Biology, École Normale Supérieure, PSL University, Paris, France, [4]Center for Computational Mathematics, Flatiron Institute, New York, NY, United States

Accurate generative statistical modeling of count data is of critical relevance for the analysis of biological datasets from high-throughput sequencing technologies. Important instances include the modeling of microbiome compositions from amplicon sequencing surveys and the analysis of cell type compositions derived from single-cell RNA sequencing. Microbial and cell type abundance data share remarkably similar statistical features, including their inherent compositionality and a natural hierarchical ordering of the individual components from taxonomic or cell lineage tree information, respectively. To this end, we introduce a Bayesian model for **t**ree-aggregated **a**mplicon and **s**ingle-**c**ell **co**mpositional **d**ata **a**nalysis (tascCODA) that seamlessly integrates hierarchical information and experimental covariate data into the generative modeling of compositional count data. By combining latent parameters based on the tree structure with spike-and-slab Lasso penalization, tascCODA can determine covariate effects across different levels of the population hierarchy in a data-driven parsimonious way. In the context of differential abundance testing, we validate tascCODA's excellent performance on a comprehensive set of synthetic benchmark scenarios. Our analyses on human single-cell RNA-seq data from ulcerative colitis patients and amplicon data from patients with irritable bowel syndrome, respectively, identified aggregated cell type and taxon compositional changes that were more predictive and parsimonious than those proposed by other schemes. We posit that tascCODA[1] constitutes a valuable addition to the growing statistical toolbox for generative modeling and analysis of compositional changes in microbial or cell population data.

Keywords: bayesian modeling, dirichlet multinomial, microbiome data, single-cell data, spike-and-slab lasso, tree aggregation, differential abundance testing

OPEN ACCESS

Edited by:
Himel Mallick,
Merck, United States

Reviewed by:
Boyu Ren,
Dana–Farber Cancer Institute, United States
Thomas P. Quinn,
Deakin University, Australia
Siyuan Ma,
University of Pennsylvania, United States

***Correspondence:**
Christian L. Müller
christian.mueller@helmholtz-muenchen.de

***Present Address:**
Salomé Carcy,
Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, NY, United States

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *29 August 2021*
Accepted: *01 November 2021*
Published: *07 December 2021*

Citation:
Ostner J, Carcy S and Müller CL (2021) tascCODA: Bayesian Tree-Aggregated Analysis of Compositional Amplicon and Single-Cell Data. Front. Genet. 12:766405. doi: 10.3389/fgene.2021.766405

1 INTRODUCTION

Next-generation sequencing (NGS) technologies have fundamentally transformed our ability to quantitatively measure the molecular make-up of single cells (Shalek et al., 2013), tissues (Regev et al., 2017; Karlsson et al., 2021), organs (He et al., 2020), as well as microbiome compositions in and on the human body (Human Microbiome Project Consortium, 2012). Single-cell RNA

[1]Available at https://github.com/bio-datascience/tascCODA.

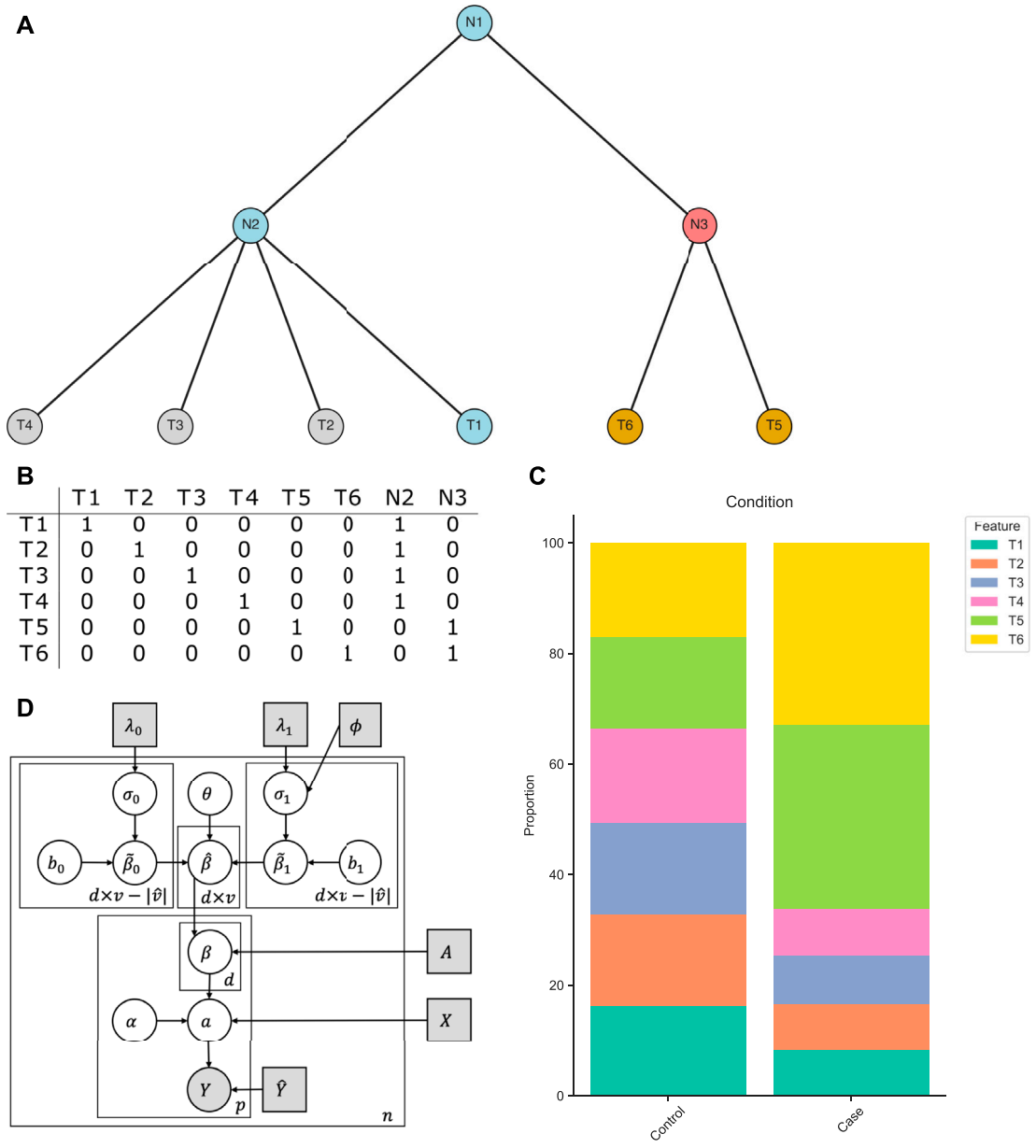
sequencing (scRNA-seq) (Tang et al., 2009; Shalek et al., 2013; Macosko et al., 2015) has become the key technology for recording the transcriptional profiles of individual cells across different tissue types (Regev et al., 2017) and developmental stages (Griffiths et al., 2018), and for determining cell type states and overall cell type compositions (Trapnell, 2015). Cell type compositions provide informative and interpretable representations of the noisy high-dimensional scRNA-seq data and are typically derived from clustering characteristic gene expression patterns in each cell (Duò et al., 2018; Traag et al., 2019), followed by analysis of the expression levels of marker genes (Luecken and Theis, 2019). As a by-product, these workflows also yield a hierarchical grouping of the cell types, either derived from the clustering procedure or determined by known cell lineage hierarchies. Determining changes in cell type populations across conditions can give valuable insight into the effects of drug treatment (Tsoucas et al., 2019) and disease status (Smillie et al., 2019), among others.

Complementary to scRNA-seq data collection, amplicon or marker-gene sequencing techniques provide abundance information of microbes across human body sites (Human Microbiome Project Consortium, 2012; Lloyd-Price et al., 2017; McDonald et al., 2018). Current estimates suggest that the human microbiome, i.e., the collection of microbes in and on the human body, outnumber an individual's somatic and germ cells by a factor of 1.3–10 (Turnbaugh et al., 2007; Sender et al., 2016). Starting from the raw read counts, amplicon data are typically summarized in count abundance tables of operational taxonomic units (OTUs) at a fixed sequence similarity level or, alternatively, of denoised amplicon sequence variants (ASVs). The marker genes also allow taxonomic classification and phylogenetic tree estimation, thus inducing a hierarchical grouping of the taxa. To reduce the dimensionality of the data set and guard against noisy and low count measurements, the taxonomic grouping information is often used to aggregate the data at a fixed taxonomic rank, e.g., the genus or family rank. Shifts in the population structure of taxa have been implicated in the host's health and have been associated with various diseases and symptoms, including immune-mediated diseases (Round and Palm, 2018), Crohn's disease (Gevers et al., 2014), and Irritable Bowel Syndrome (IBS) (Ford et al., 2017).

In the present work, we exploit the remarkable similarities between scRNA-seq-derived cell type data and amplicon-based microbial count data and propose a statistical generative model that is applicable to both data modalities: the Bayesian model for **t**ree-aggregated **a**mplicon and **s**ingle-**c**ell **CO**mpositional **D**ata **A**nalysis, in short, tascCODA. Our model assumes that count data are available in the form of a $n \times p$-dimensional count matrix Y containing the counts of p different cell types or microbial taxa in n samples, a covariate matrix $n \times d$-dimensional X carrying metadata or covariate information for each sample, and a tree structure with p leaves that imposes a hierarchical order on the count data Y. Since both amplicon and scRNA-seq technologies are limited in the amount of material that can be processed in one sample, the total number of counts in rows of Y do not reflect total abundance measurements of the features but rather relate to the efficiency of the sequencing experiment itself (Gloor et al., 2017). This implies that the counts only carry relative abundance information, making them essentially compositional data (Aitchison, 1982).

tascCODA is a fully Bayesian model for tree-aggregated modeling of count data and is a natural extension of the scCODA model, recently introduced for compositional scRNA-seq data analysis (Büttner et al., 2020). At its core, tascCODA models the count data Y via a Dirichlet Multinomial distribution and associates count data and covariate information via a log-link function. To encourage sparsity in the underlying associations between the covariates and the hierarchically grouped features, tascCODA exploits recent ideas from tree-guided regularization and the spike-and-slab LASSO (Ročková and George (2018)). This allows tascCODA to perform tree-guided sparse regression on compositional responses with any type or number of covariates. In particular, in the presence of a single binary covariate, e.g., a condition indicator, tascCODA allows to perform Bayesian differential abundance testing. More generally, however, tascCODA enables to determine how host phenotype, such as disease status, host covariates such as age, gender, or an individual's demographics, or environmental factors jointly influence the compositional counts. Finally, incorporating tree information into the inference allows tascCODA to not only identify associations between individual features, but also entire groups of features that form a subset of the tree.

tascCODA complements several recent statistical approaches, in particular, from the field of microbiome data analysis, some of which also use the concept of tree-guided models. Chen and Li (2013) were among the first to use the sparse Dirichlet-Multinomial model to connect compositional count data with covariate information in a penalized maximum-likelihood setting. Wadsworth et al. (2017) were the first to use a similar model in a Bayesian setting. Both adaANCOM (Zhou C. et al. (2021)) and the Logstic-tree normal model (Wang et al. (2021)) use the Dirichlet-tree (multinomial) model (Wang and Zhao (2017)) to determine differential abundance of microbial taxa via a product of Dirichlet distributions at each split. The PhILR model (Silverman et al., 2017) uses the phylogenetic tree of a microbial community to compute an isometric logratio transform with interpretable balances. Furthermore, there are recent advances in constructing optimal hierarchical partitions of HTS data and to predict variables of interest from them (Quinn and Erb, 2019; Gordon-Rodriguez et al., 2021), that do not rely on pre-defined trees, but rather structure the data in the best way to be predictive of the outcome. These methods restrict themselves, however, to fully binary trees. On the other hand, the trac method (Bien et al., 2021) uses tree-guided regularization (Yan and Bien, 2021) in a maximum-likelihood-type framework to predict continuous outcomes from compositional microbiome data.



	T1	T2	T3	T4	T5	T6	N2	N3
T1	1	0	0	0	0	0	1	0
T2	0	1	0	0	0	0	1	0
T3	0	0	1	0	0	0	1	0
T4	0	0	0	1	0	0	1	0
T5	0	0	0	0	1	0	0	1
T6	0	0	0	0	0	1	0	1

FIGURE 1 | Intuition behind tascCODA. **(A)** A multifurcating tree structure $\mathcal{T}$ with internal nodes N1, N2, N3, and tips T1 . . .T6. tascCODA decides whether modeling the change of abundance of a subtree (e.g. nodes T5, T6 - gold). as a common effect at their common ancestor (e.g., N3 - red) is preferable. The blue nodes T1, N1, and N2 are reference nodes in this example. **(B)** Ancestor matrix of the tree in **(A)**. **(C)** Example dataset where the abundances of T5 and T6 increase in the same way between conditions (relative to the reference T1). Here, a group-level effect on N3 would be the preferred option. **(D)** Plate representation of the tascCODA model. Grey squares indicate fixed parameters and input variables that are either part of or directly calculated from the data. The grey circle represents the output count matrix, white circles show latent variables.

In its present form, the Bayesian model behind tascCODA is ideally suited for data sets of moderate dimensionality, typically $p < 100$, yet can handle extremely small sample sizes n. Since amplicon datasets are usually high-dimensional in the number of taxa and exhibit high overdispersion and excess number of zeros, we focus on the analysis of genus-level microbiome data. In the context of cell type compositional data, on the other hand, often only very few replicate samples are available (Büttner et al., 2020). Here, tascCODA can leverage well-calibrated prior information to operate in low-sample regimes where frequentist methods likely fail.

The remainder of the paper is structured as follows. In the next section, we introduce the tascCODA model and describe the computational implementation. In **Section 3**, we describe and discuss synthetic data benchmarks and provide two real-world applications, on human single-cell RNA-seq data from ulcerative

colitis patients and amplicon data from patients with irritable bowel syndrome. Finally, we summarize the key points in **Section 4** and present considerations about future extensions of the method. A flexible and user-friendly implementation of tascCODA is available in the Python package *tascCODA*[2]. All results in this paper are fully reproducible and available on Zenodo[3].

2 MATERIALS AND METHODS

2.1 Model Description

We start with formally describing the problem at hand. Let $Y \in \mathbb{R}^{n \times p}$ be a count matrix describing n samples from p features (e.g., cell types, microbial taxa, etc.), and $X \in \mathbb{R}^{n \times d}$ be a matrix that contains the values of d covariates of interest for each sample. Due to the technical limitations of the sampling procedure, the sum of counts in each sample, $\bar{Y}_i = \sum_{j=1}^{p} Y_{i,j}$ must be seen as a scaling factor, making the data compositional (Gloor et al. (2017)). Additionally, the features described by Y are hierarchically ordered by a tree $\mathcal{T}$ with p leaves and t internal nodes, resulting in a total number of $v = p + t$ nodes in $\mathcal{T}$ (**Figure 1A**). Such tree structures are usually motivated by taxonomy (McDonald et al., 2012; Quast et al., 2013), determined by phylogenetic similarities (Schliep, 2010), or obtained via serial binary partitions (Quinn and Erb, 2019). The tree can further be bifurcating or multifurcating, thus internal nodes may have two or more descendants.

$\mathcal{T}$ can be fully characterized by a binary ancestor matrix $A \in \{0,1\}^{p \times v}$. Hereby, each row of A stands for a feature or leaf node of $\mathcal{T}$, the first p columns also denote the leaves of the tree, and the last t columns represent the internal nodes. The entries $A_{j,k}$ are 1, if column k corresponds either to feature j ($j = k$) or to one of its parents, otherwise it is 0 (**Figure 1B**):

$$A_{j,k} = \begin{cases} 1 & \text{if } j = k \text{ or } k \text{ is ancestor of } j \\ 0 & \text{else.} \end{cases}$$

Our goal is to determine how changes in abundance of features (leaves of $\mathcal{T}$) are associated with the covariates in X, and select a sparse set of the most important covariate-feature effects. To achieve an even more parsimonious result, we further determine whether groups of features that form subtrees of $\mathcal{T}$ are affected by the conditions in the same manner (**Figure 1A**), and model them with a common effect if possible. This group-wise modeling step not only gives an accurate, yet easy to interpret description of the changes in the feature composition, but can also reveal shared traits among structural subgroups of features that might be missed in analyses that do not take the tree structure into account.

2.1.1 Core Model With Tree Aggregation

tascCODA posits a Dirichlet-Multinomial model for $Y_{i,\cdot}$ for each sample $i \in 1\ldots, n$, thus accounting for the compositional nature of the count data. The covariates are associated with the features through a log-linear relationship. We put uninformative Normal priors on the base composition α, which describes the data in the case $X_{i,\cdot} = 0$:

$$Y_i \sim \text{DirMult}\left(\bar{Y}_i, \mathbf{a}(X)_i\right) \tag{1}$$

$$\log\left(\mathbf{a}(X)\right)_i = \alpha + X_{i,\cdot}\beta \tag{2}$$

$$\alpha_j \sim \mathcal{N}(0, 10) \qquad \forall j \in [p]. \tag{3}$$

The total count $\bar{Y}_i$ is directly inferred from the data for each sample. The effect of the lth covariate on the jth feature is therefore given by $\beta_{l,j}$.

We now use a variant of the tree-based penalty formulation of Yan and Bien (2021) to model common effects at each internal node of $\mathcal{T}$ in addition to the effects on the leaves. We define a node effect matrix $\hat{\beta} \in \mathbb{R}^{d \times v}$ and associate aggregations on internal nodes with the correct tips by multiplying with the ancestor matrix A:

$$\beta = \hat{\beta} A^T \tag{4}$$

To illustrate the intuition behind this step, we consider an example based on the tree in **Figure 1A**. In a binary covariate setting, the features T1-T6 are uniformly distributed in the control population, while in the case population, the abundance of features T5 and T6 (with respect to feature T1) is greatly increased by the same relative amount (**Figure 1C**). Instead of having two equally-sized effects on the components of $\hat{\beta}$ corresponding to T5 and T6, the same can be achieved in tascCODA with only one parameter by placing an effect on the internal node N3. Through **Eq. 4**, this effect is propagated to the leaves T5 and T6 in β in order to model the population.

While this aggregation step can significantly reduce the number of parameters needed to describe the changes in the data, the solution is not unique. An effect on an internal node is equivalent to effects of the same size on all its descendant leaves. Therefore, the number of nonzero entries in $\hat{\beta}$ must be controlled, raising the need for a sparse selection of the most important effects. While in the example above, the reduction of nonzero effects by using a group aggregation on node N3 clearly outweighs the loss in accuracy by assuming that features T5 and T6 behave in the same manner, this trade-off might not be as clear in real datasets. We thus also need a way to adjust the model towards selecting either more sparse and generalizing, or more detailed and less parsimonious solutions.

2.1.2 Spike-And-Slab Lasso Prior

To ease model interpretability, many statistical models provide a mechanism for obtaining sparse model solutions. In high-dimensional linear regression, this can be achieved via the lasso (Tibshirani, 1996), which adds an $\mathcal{L}_1$-penalty on the regression coefficients. In Bayesian modeling, spike-and-slab priors are a popular choice to perform automatic model selection. Recently, Ročková and George (2018), developed a connection between the two approaches in the form of the spike-and-slab lasso prior, which provides a Bayesian equivalent to penalized likelihood estimation. The spike-and-

[2]https://github.com/bio-datascience/tascCODA.
[3]https://zenodo.org/record/5302136.

slab lasso prior describes each component of $\hat{\beta}_{l,k}$ as a mixture of two double-exponential priors with different rates $\lambda_{0,l,k}$, $\lambda_{1,l,k}$ and a shared mixture coefficient θ:

$$\hat{\beta}_{l,k} = \theta\tilde{\beta}_{1,l,k} + (1-\theta)\tilde{\beta}_{0,l,k} \qquad \forall k \in [\nu], l \in [d] \quad (5)$$

$$\tilde{\beta}_{m,l,k} = \sigma_{m,l,k} * b_{m,l,k} \qquad \forall k \in [\nu], m \in \{0,1\}, l \in [d] \quad (6)$$

$$\sigma_{m,l,k} \sim \mathrm{Exp}\left(\lambda_{m,l,k}^2/2\right) \qquad \forall k \in [\nu], m \in \{0,1\}, l \in [d] \quad (7)$$

$$b_{m,l,k} \sim \mathcal{N}(0,1) \qquad \forall k \in [\nu], m \in \{0,1\}, l \in [d] \quad (8)$$

$$\theta \sim \mathrm{Beta}(1, 1/\nu) \quad (9)$$

This prior can be reformulated as a likelihood penalty function that represents a combination of weak penalization of larger effects by $\lambda_{1,l,k}$ and strong penalization of effects close to zero by $\lambda_{0,l,k}$, respectively (See **Supplementary Material Section 1.2**). As recommended by Ročková and George (2018), we use the non-separable version of the spike-and-slab lasso prior, which provides self-adaptivity of the sparsity level and an automatic control for multiplicity via a Beta prior on θ (Bai et al. (2020a); Scott and Berger (2010)). We further set $\lambda_{0,l,k} = 50\ \forall l, k$ to achieve a strong penalization in the "spike" part of the prior, leaving $\lambda_{1,l,k}$ as our only parameter that controls the total amount of penalty applied at larger effect values.

2.1.3 Node-Adaptive Penalization

We use a variant of the strategy proposed by Bien et al. (2021) to make the strength of the regularization penalty dependent on the corresponding node's position in the tree. We introduce the following sigmoidal scaling:

$$\lambda_{1,l,k} = 2\lambda_1 \frac{1}{1 + e^{-\phi\left(L_k/p - 0.5\right)}} \ \forall l, \quad (10)$$

where $\lambda_1 = 5$ is the default value for the penalty strength, L_k is the number of leaves that are contained in the subtree of node k, and ϕ acts as a scaling factor based on the tree structure. If $\phi = 0$, the default in tascCODA, all nodes are penalized equally with λ_1, while for $\phi < 0$, effects on nodes with larger subtrees, located closer to the root of the tree, are penalized less and are therefore more likely to be included in the model. If $\phi > 0$, a solution that comprises more diverse effects on leaf nodes will be preferred. Thus, the parameter ϕ provides a way to trade off model accuracy with the level of aggregation. We discuss the behavior of the spike-and-slab LASSO penalty and the choice of $\lambda_{0,1}$ in more detail in the **Supplementary Material**.

2.1.4 Reference Feature

Since the data at hand is compositional, model uniqueness and interpretability are only guaranteed with respect to a reference. Popular choices include picking one of the p features or the (geometric) mean over multiple or all groups (Fernandes et al., 2014). Following the scCODA model, we pick a single reference feature prior to analysis (Büttner et al., 2020). Technically, this is achieved by choosing one feature $\hat{p}$ that is set to be unchanged by all covariates. Let $\hat{\nu}$ be the set of ancestors of $\hat{p}$. By forcing $\hat{\beta}_{l,k} = 0\ \forall k \in \hat{\nu}, l \in [d]$, we ensure that the reference is not influenced by the covariates through any of its ancestor nodes. If no suitable reference feature is known a priori, tascCODA provides an automatic way of selecting the feature with minimal dispersion across all samples among the features that are present in at least a share of samples t (default $t = 0.95$; this value can be lowered if no suitable feature exists).

$$\hat{p} = \arg\min_{j=1,\ldots,p} \mathrm{Disp}\left(Y'_{\cdot,j}\right) \ s.t.\ |i\colon Y_{i,j} > 0|/n \geq t$$

The restriction to large presence avoids choosing a rare feature as the reference where small changes in terms of counts lead to large relative deviations. The least-dispersion approach is aimed at reducing the bias introduced by the choice of reference. **Eqs. 1–9** together with the reference feature yields the tascCODA model (**Figure 1D**):

$$\begin{aligned}
Y_i &\sim \mathrm{DirMult}\left(\bar{Y}_i, \mathbf{a}(X)_i\right) \\
\log\left(\mathbf{a}(X)\right)_i &= \boldsymbol{\alpha} + X_{i,\cdot}\beta \\
\alpha_j &\sim \mathcal{N}(0,10) && \forall j \in [p] \\
\beta &= \hat{\beta}A^T \\
\hat{\beta}_{l,k} &= 0 && \forall k \in \hat{\nu}, l \in [d] \\
\hat{\beta}_{l,k} &= \theta\tilde{\beta}_{1,l,k} + (1-\theta)\tilde{\beta}_{0,l,k} && \forall k \in \{[\nu]\backslash\hat{\nu}\}, l \in [d] \\
\tilde{\beta}_{m,l,k} &= \sigma_{m,l,k} * b_{m,l,k} && \forall k \in \{[\nu]\backslash\hat{\nu}\}, m \in \{0,1\}, l \in [d] \\
\sigma_{m,l,k} &\sim \mathrm{Exp}\left(\lambda_{m,l,k}^2/2\right) && \forall k \in \{[\nu]\backslash\hat{\nu}\}, l \in \{0,1\}, l \in [d] \\
b_{m,l,k} &\sim \mathcal{N}(0,1) && \forall k \in \{[\nu]\backslash\hat{\nu}\}, l \in \{0,1\}, l \in [d] \\
\theta &\sim \mathrm{Beta}\left(1, \frac{1}{|\{[\nu]\backslash\hat{\nu}\}|}\right)
\end{aligned}$$

with the default choices of $\lambda_{0,l,k} = 50$ and $\lambda_{1,l,k}$ set according to (10) with hyperparameters ϕ and $\lambda_1 = 5$ (**Supplementary Material Section 1.2**).

2.2 Computational Aspects

Before performing Bayesian inference with the tascCODA model, several data preprocessing steps are applied. Singular nodes, i.e., internal nodes that have only one child node, are removed from the tree, since their effect only propagates to one node and is therefore redundant. We also add a small pseudo-count of 0.5 to all zero entries of Y to minimize the frequency of numerical instabilities in our tests. Finally, we recommend normalizing all covariates to a common scale before applying tascCODA to avoid biasing the model selection process toward the covariate with the largest range of values.

Because tascCODA is a hierarchical Bayesian model, we use Hamiltonian Monte Carlo sampling (Betancourt and Girolami, 2015) for posterior inference, implemented through the tensorflow (Abadi et al., 2016) and tensorflow-probability (Dillon et al., 2017) libraries for Python, solving the gradient in each step via automatic differentiation. By default, tascCODA uses a leapfrog integrator with Dual-averaging step size adaptation (Nesterov, 2009) and 10 leapfrog steps per iteration, sampling a chain of 20,000 posterior realizations and discarding the first 5,000 iterations as burn-in, which was also the setting for all applications in this article, unless explicitly stated otherwise. As an alternative, No-U-turn sampling (Homan and

Gelman, 2014) is available for use with tascCODA. The initial states for all α_j and $b_{m,l,k}$ are randomly sampled from a standard normal distribution. All $\sigma_{m,l,k}$ and θ values are initialized at 1 and 0.5, respectively.

To determine the credible effects of covariates on nodes from the chain of posterior samples, we calculate the threshold of practical significance δ_k, introduced by Ročková and George (2018), for each node:

$$\delta_k = \frac{1}{\lambda_0 - \lambda_{1,k} \log\left(\frac{1}{p^*_{\theta,k}(0)} - 1\right)} \tag{11}$$

$$p^*_{\theta,k}(\beta) = \frac{\theta^* \frac{\lambda_{1,k}}{2} e^{-\lambda_{1,k}|\beta|}}{\theta^* \frac{\lambda_{1,k}}{2} e^{-\lambda_{1,k}|\beta|} + (1-\theta^*)\frac{\lambda_0}{2} e^{-\lambda_0|\beta|}} \tag{12}$$

Here, $\theta^\star$ is the posterior median of θ. More details on δ are available in the **Supplementary Material**. We compare the posterior median effects $\hat{\beta}^*_{l,k}$ to the corresponding δk and select all effects where $|\hat{\beta}^*_{l,k}| > \delta_k$ as credible, otherwise they will be set to 0, resulting in $\hat{\beta}^{(C)}$, the matrix with only credible effects,

$$\hat{\beta}^{(C)}_{l,k} = \begin{cases} \hat{\beta}^*_{l,k} & \text{if } |\hat{\beta}^*_{l,k}| > \delta_k \\ 0 & \text{else.} \end{cases} \tag{13}$$

In most applications, the nonzero entries of $\hat{\beta}^{(C)}$ are of primary interest, which directly show how the covariates influence sets of features defined by the tree structure. Their sign indicates whether the effect corresponds to an increase ($\hat{\beta}^{(C)}_{l,k} > 0$) or a decrease ($\hat{\beta}^{(C)}_{l,k} < 0$). Due to the compositional data properties introduced by the Dirichlet-Multinomial, its expectation

$$\mathrm{E}\left[Y_i \sim \mathrm{DirMult}\left(\bar{Y}_i, \mathbf{a}(\mathbf{x})_i\right)\right] = \bar{Y}_i \frac{\mathbf{a}(\mathbf{x})_i)}{\sum_{j=1}^{p} \mathbf{a}(\mathbf{x})_i)_j} \tag{14}$$

can not be separated by the individual features. Because the shifts in $\mathrm{E}[Y_i]$ caused by effects $\hat{\beta}$ are dependent on the total sum $\sum_{j=1}^{p} e^{\alpha_j + X(\hat{\beta}A^T)_j}$ through **Eqs. 2**, **4**, **14**, a credible effect on any feature or aggregation has an impact on the posterior mean counts of all features, i.e. a relative increase in one feature will also induce a decrease of all other features (Gloor et al., 2017). Therefore, a quantitative interpretation of effect sizes is only possible in a limited sense. Within the same model, larger changes will correspond to larger absolute values $|\hat{\beta}_{l,k}|$, but they are not comparable across multiple runs of tascCODA.

In the context of differential abundance testing, we can additionally obtain the set of differentially abundant features D by multiplying $\hat{\beta}^{(C)}$ with A^T, and get

$$D = \left\{(l, j) \in [d] \times [p] : \left(\hat{\beta}^{(C)}_{l,k} A^T\right)_j \neq 0\right\} \tag{15}$$

as the set of features that are part of at least one credible effect.

A Python package for tascCODA is available at https://github.com/bio-datascience/tascCODA. Building upon the scCODA package, the software provides methods to seamlessly integrate scRNA-seq data from scanpy (Wolf et al., 2018) or microbial population data via pandas (McKinney, 2010). The package also allows to perform differential abundance testing with tascCODA and visualize tascCODA's results through tree plots from the toytree package. All results were obtained using Python 3.8 with tensorflow = 2.5.0 (Abadi et al. (2016)), tensorflow-probability = 0.13 (Dillon et al. (2017)), arviz = 0.11 (Kumar et al. (2019)), numpy = 1.19.5, scanpy = 1.8.1 (Wolf et al. (2018)), toytree = 2.0.1, and sccoda = 0.1.4 (Büttner et al. (2020)).

3 RESULTS

3.1 Simulation Studies

3.1.1 Model Comparison

To test the performance of tascCODA in a differential abundance testing scenario, we generated compositional datasets with an underlying tree structure and compared how well several models could detect the changes introduced by a binary covariate. For compositional models that do not account for the tree structure, we used the state-of-the art methods ANCOM-BC (Lin and Peddada (2020)), ANCOM (Mandal et al. (2015)), and ALDEx2 (Fernandes et al. (2014)) from the field of microbiome data analysis, as well as scCODA (Büttner et al., 2020) from scRNA-seq analysis. Based on the recommendations by Aitchison (1982), we also analyzed the data with the additive log-ratio (ALR) transformation in combination with t- or Wilcoxon rank-sum tests. We also included the recent adaANCOM (Zhou C. et al., 2021), a differential abundance testing method that accounts for the tree structure. Furthermore, we applied tascCODA with different values for the aggregation parameter, ϕ = (−10, −5, −1, 0, 1, 5, 10), setting $\lambda_1 = 5$.

We first defined four different data sizes p = (10, 30, 50, 100) and randomly generated a multifurcating tree with depth five for each value of p. We then chose three nodes (one internal on the level directly above the leaves, two leaves) from each tree, whose child leaves, denoted by p', are set to be differentially abundant under a binary (control-treatment) condition (**Supplementary Figures S2–S5**). Similar to Wadsworth et al. (2017), we generated $n = n_0 + n_1$ compositional data samples from two groups of equal size $n_0 = n_1$ = (5, 20, 30, 50). Each sample Y_i is a realization of a Dirichlet-Multinomial distribution with a total sum of $\bar{Y}_i$ = 10, 000 and a parameter vector $\gamma^\star$. For extra dispersion in the data, we set $\gamma^*_i = \frac{\gamma_i}{\sum_j \gamma_j} \frac{1-\psi}{\psi}$ with $\psi = 0.002$. The parameters for the first (control) group were generated via $\gamma_{0,i} = \exp(\alpha_i)$; α_i ~Unif(−2, 2). In the second (treatment) group, we added an effect β = (0.3, 0.5, 0.7, 0.9) to the components in p': $\gamma_{1,i} = \exp(\alpha_i + \beta \mathbb{I}_{(i \in p')})$. For each parameter combination (p, n_0, β), we randomly generated 20 replicates, resulting in a total of 1280 datasets.

Since the adaANCOM method assumes a bifurcating tree structure, we transformed each tree node to a series of bifurcating splits via the *multi2di* and *collapse.singles* methods from the *ape* package for R (Paradis et al. (2004)) before applying the method. For the methods that require a reference category (ALR, scCODA, tascCODA, ALDEx2), we used the last component, which was always designed to be unaffected by

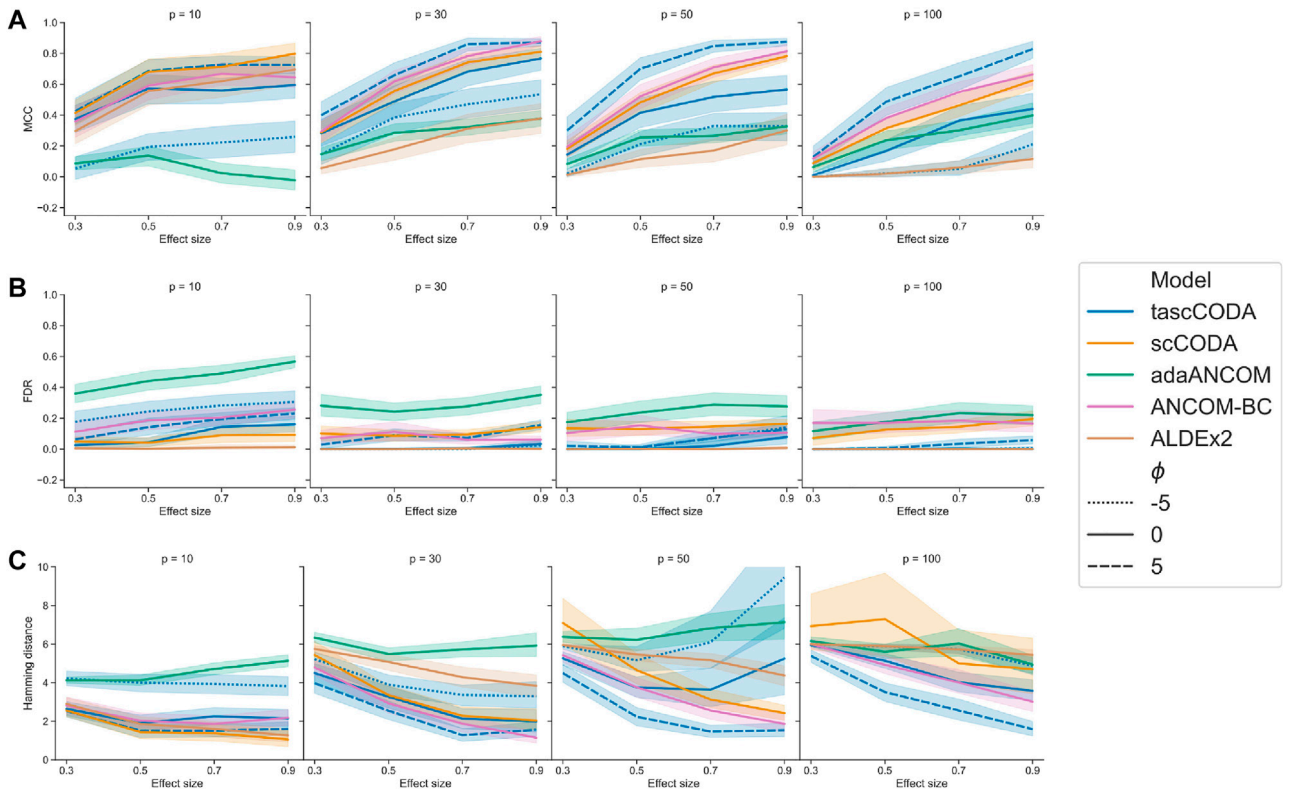


FIGURE 2 | Performance comparison of tascCODA and other methods on simulated data with one binary covariate (differential abundance testing). Plots are grouped by the number of simulated components p and the effect size β. For tascCODA, different values of ϕ were tested (dashed blue lines). The areas around each line represent the standard deviation. Performance measured by **(A)** Matthews correlation coefficient (MCC). **(B)** False discovery rate (FDR) **(C)** Hamming distance between ground truth and determined effects.

the condition, as the reference. After applying each method to a dataset, we corrected the resulting p-values by the Benjamini-Hochberg procedure, where applicable, except for ANCOM-BC, where we used the recommended Holm correction of p-values, and determined the significant results at an expected FDR level of 0.05. The Bayesian methods scCODA and tascCODA do not produce p-values and identify credible effects as previously described.

For an overall indicator of how well the different methods could determine differentially abundant features, we considered Matthews correlation coefficient (**Figure 2A**). Here, adaANCOM showed poor performance especially on small datasets, while ALDEx2 struggled when p was larger. Only scCODA and ANCOM-BC performed well in comparison for all data and effect sizes. For tascCODA, varying the aggregation level ϕ had a strong influence on the performance. With larger values of ϕ, tascCODA prefers less generalizing effects, resulting in a more detailed solution and larger MCC. At a high resolution level ($\phi = 5$), tascCODA was on par with or even better than scCODA and ANCOM-BC, showing almost no sensitivity to the size of the dataset. Because the trees in our simulation contained only effects on leaf nodes or the level directly above, preferring generalizing effects ($\phi = -5$) resulted in worse performance, while the unbiased case of $\phi = 0$ gave slightly worse results than scCODA and ANCOM-BC. All methods shown in **Figure 2B** except adaANCOM controlled the FDR reasonably well, although ANCOM-BC and scCODA could not always hold the nominal level of 0.05. Only ALDEx2, which is known to be very conservative (Hawinkel et al., 2019; Büttner et al., 2020), produced almost no false positives, at the cost of larger type 2 error. tascCODA had a slightly inflated FDR (< 0.25) for smaller values of ϕ in some cases, which became more apparent when analyzing the ability of each method to exactly recover the true effects (**Figure 2C**). Increasing the effect size resulted in a reduced Hamming distance between the ground truth and tascCODA with $\phi = 5$, which consistently outperformed all other models. tascCODA in the misspecified setting $\phi = -5$ showed an inflated Hamming distance, especially for $p = 30$. This is, however, expected since tascCODA is forced to infer small-sized effects at the top level, resulting in many falsely detected features and thus a large deviation from the true sparse solution. In practice, this highlights the need to perform cross-validation over different levels of ϕ to reduce false discoveries due to misspecification. We further found that ANCOM detected many false positives in all of our simulations, while the ALR-based methods were similarly

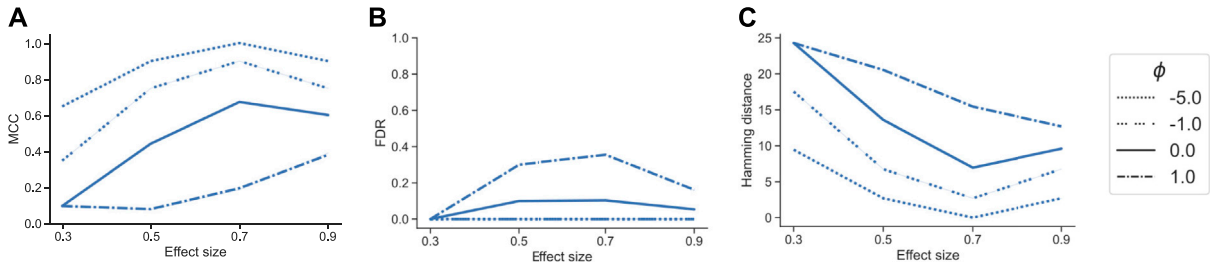


FIGURE 3 | Performance comparison of different bias settings for tascCODA on simulated data with the effect being located near the root of the tree, depending on effect size. Performance measured by **(A)** Matthews correlation coefficient (MCC). **(B)** False discovery rate (FDR) **(C)** Hamming distance between ground truth and determined effects.

conservative as ALDEx2 (**Supplementary Figures S8–S10**). Increasing the sample size generally improved the recovery performance of all methods except for tascCODA with misspecified ϕ (**Supplementary Figure S10**).

3.1.2 Effect Detection at High Tree Levels

In the next benchmark scenario, we evaluated the effect of the tuning parameter ϕ in tascCODA to detect effects on larger groups of features through aggregation at higher levels of the tree. To this end, we considered the $p = 30$ setting with the tree structure from **Supplementary Figure S5**, and defined an effect on a node near the root, influencing almost all features (**Supplementary Figure S6**). We simulated datasets in the same manner as for the previous benchmark, with $n = 10$, $\beta = (0.3, 0.5, 0.7, 0.9)$, and 20 replicates per effect size. We then compared tascCODA with different levels of ϕ using the same performance metrics as before.

With a correctly specified parametrization $\phi < 0$, favoring effects near the root, tascCODA recovered almost all relevant effects, as indicated by a small Hamming distance and high MCC, without producing false positive results (**Figure 3**). With increasing ϕ, however, tascCODA favors effects on the leaves, thus entering the misspecified regime. As predicted, tascCODA was able to only recover a small portion of the true effects, while producing more false positive results. This highlights tascCODA's ability to consistently uncover effects on larger groups of features which would be missed when not taking into account tree information.

3.1.3 Simulation With Multiple Covariates

In our third benchmark scenario, we simulated data with two covariates to showcase how tascCODA is able to distinguish effects from two different sources. Taking the tree from the method comparison study with $p = 30$ (**Supplementary Figure S3**), we first defined a binary covariate x_0 with effect sizes $\beta_0 = (0.3, 0.5, 0.7, 0.9)$ as before, and $n = 10$ samples per group. We also included a second covariate $x_1 \sim Unif(0, 1)$ with effect size $\beta_1 = 3$ that affects node 39 and therefore features 13–23 in all samples. For each effect size, we simulated 10 datasets and applied tascCODA with $\phi = (-5, 0, 5)$ and two different design matrices X. For the first design matrix, we used only x_0, while the second design matrix contained both x_0 and x_1 as covariates. We compared how well both configurations could recover the effects introduced by x_0 in terms of MCC, FDR, and Hamming distance to the ground truth.

Ignoring x_1 in the model design resulted in an overall worse performance of tascCODA for all metrics, all effect sizes for x_0, and all values of ϕ (**Figure 4**). In every case it proved beneficial to include the second covariate in the model, resulting in almost no false positive detections of changes caused by the first covariate. Further, the two-covariate model achieved an MCC and Hamming distance that were similar to our simulations where only one covariate acted on the data (**Figure 2**). This proves that tascCODA is able to reliably identify the influence of multiple covariates on the count data.

3.2 Experimental Data Applications

3.2.1 Single-cell Sequencing Analysis of Ulcerative Colitis in Humans

Ulcerative colitis is one of the most common manifestations of inflammatory bowel disease. The disease alternates between periods of symptomatic flares and remissions. The flares are due to the surge of an inflammatory reaction in the colon, causing superficial to profound ulcerations, which manifests with bloody stool, diarrhea and abdominal pain. The patients will thus have part of their colon referred to as "inflamed", while colonic tissue still seemingly intact will be called "non-inflamed". To show how tascCODA can be applied to cell population data from scRNA-seq experiments, we used data collected by Smillie et al. (2019) from a study of the colonic epithelium on ulcerative colitis (UC). In the study, a total of 133 samples from 12 healthy donors, as well as inflamed and non-inflamed tissue from 18 patients with UC, were obtained via single-cell RNA-sequencing, divided into epithelial samples and samples from the Lamina Propria (**Supplementary Data 1.3.1**).

We applied tascCODA to six different subsets of the data, comparing two of the three health conditions in one type of tissue at a time, and then compared our findings with the results of scCODA and the Dirichlet regression model used by Smillie et al. (2019), implemented in the *DirichletReg* package for R (Maier (2014)). For tascCODA and scCODA, we used the automatically determined reference cell types, which are identical for both models in all cases, and applied scCODA

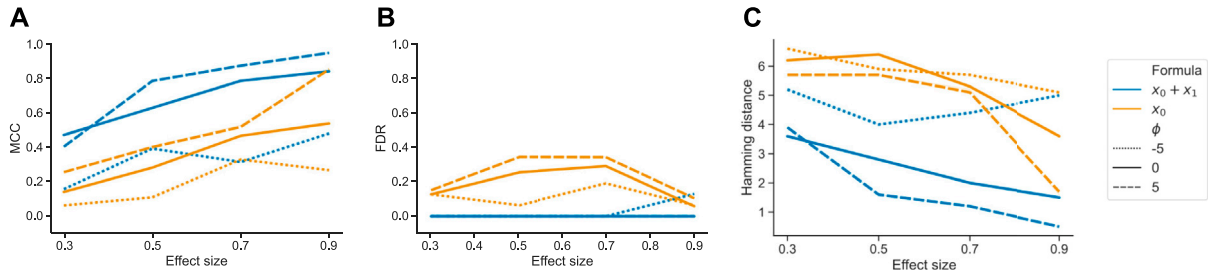


FIGURE 4 | Performance comparison for tascCODA on simulated data with two covariates. The setups including both or only one covariate in the model are shown as $x_0 + x_1$ and x_0, respectively. Simulations were evaluated for different effect sizes and aggregation levels ϕ. Performance measured by **(A)** Matthews correlation coefficient (MCC). **(B)** False discovery rate (FDR) **(C)** Hamming distance between ground truth and determined effects.

with an FDR level of 0.05. In the Dirichlet regression model, we adjusted the *p*-values by the Benjamini-Hochberg procedure, and selected differentially abundant cell types at a level of 0.05.

The cell lineage tree inferred from Smillie et al. (2019) is divided into epithelial, stromal and immune cells at the top level (**Figure 5**). While the biopsies from the Epithelium contain mostly epithelial cells, and samples from the Lamina Propria consist of cells mostly from the other two lineages, both groups also include considerable amounts of cells from the other major lineages. We first compared scCODA and Dirichlet regression, which both do not take the tree structure into account, to tascCODA with $\phi = 5$ (**Figure 6**), thus preferring a detailed solution with effects mainly located on leaf nodes, which approaches the leaf-only solutions of the other two methods. In this setting, tascCODA, scCODA and Dirichlet regression all determined mostly epithelial cells to shift in abundance between pairwise comparisons of healthy, non-inflamed, and inflamed tissue samples from the intestinal Epithelium (**Figure 6A**), and most changes in the Lamina Propria to be among stromal and immune cells (**Figure 6B**). When propagating the node effects of tascCODA with $\phi = 5$ to the leafs via **Eq. 15**, the differentially abundant cell types determined by tascCODA, scCODA, and Dirichlet regression were largely identical (**Figure 6**).

To further investigate the predictive and sparsity-inducing powers of tascCODA, we performed out-of-sample prediction with the results obtained from tascCODA and scCODA on 5-fold cross validation splits of each of the six data subsets. For both models, we determined cell type-specific effect vectors $\beta^\star$ (tascCODA: $\beta^* = A\hat{\beta}_j^{(C)}$, as in **Eq. 15**; scCODA: Model output) as well as the posterior mean of the base composition $\alpha^\star$ on the training splits, and used them to predict cell counts for each health status label X_l in the corresponding test split as $\hat{y}_{j,l} = \frac{e^{\alpha_j^* X_l \beta_j^*}}{\sum_{j=1}^{p} e^{\alpha_j^* X_l \beta_j^*}} \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} \bar{Y}_i$. We measured the predictive power of tascCODA and scCODA as the mean squared logarithmic error (MSLE) between the actual and predicted cell counts, and sparsity as the average number of nonzero effects over all five splits (**Table 1**). For small ϕ, tascCODA determined very few or no credible effects, while the MSLE was usually slightly higher than the MSLE from scCODA. In unbiased setting $\phi = 0$, tascCODA found credible effects in three scenarios, which considerably reduced the MSLE. With a small bias towards the leaves ($\phi = 1$), tascCODA even outperformed scCODA in terms of MSLE in one case, while for $\phi = 5$, tascCODA achieved a lower MSLE and similar number of credible effects in three scenarios, and a lower number of credible effects and similar MSLE in the other three scenarios. We observed a curious result when comparing non-inflamed and inflamed epithelial samples. Here, the MSLE increased with rising ϕ, indicating that the mean model over all samples described the data better than trying to determine variation between the two groups. This confirms the intuition that the aggregation bias ϕ in tascCODA acts as a trade-off between generalization level and prediction accuracy. For smaller ϕ, tascCODA will select fewer, more general effects, which might miss subtle changes at a lower level of the lineage tree, while with increasing ϕ, tascCODA's results will approach the ones discovered without taking tree aggregation into account.

For a more detailed comparison between tascCODA and scCODA, we compared healthy to non-inflamed biopsies of control and UC patients. When choosing $\phi = 5$, thus biasing tascCODA towards the leaf nodes, tascCODA detected the differences in cell composition in the Epithelium as changes in abundance of the same 3 cell types as scCODA (**Figure 5A**). In the Lamina Propria, tascCODA detected credible changes on six different groups of cell types, including T and B cells, which were previously linked to UC (Holmén et al. (2006); Smillie et al. (2019)), as well as eight single cell types (**Figure 5B**). Notably, tascCODA amplified the decrease of Plasma B-cells induced by the group effect on B-cells by an additional negative effect on the cell type level. A strong decrease of Plasma cells was also confirmed by Smillie et al. (2019) through FACS stainings. Importantly, tascCODA described the data with only 14 nonzero effects, whereas with scCODA, 21 credible effects were produced.

As a contrast, we also examined the unbiased setting with $\phi = 0$, treating all nodes equally. Here, the cell type-specific changes in the Epithelium were not picked up anymore by tascCODA (**Figure 5C**). In the Lamina Propria, only seven effects, almost all on groups of cell types, were detected by tascCODA

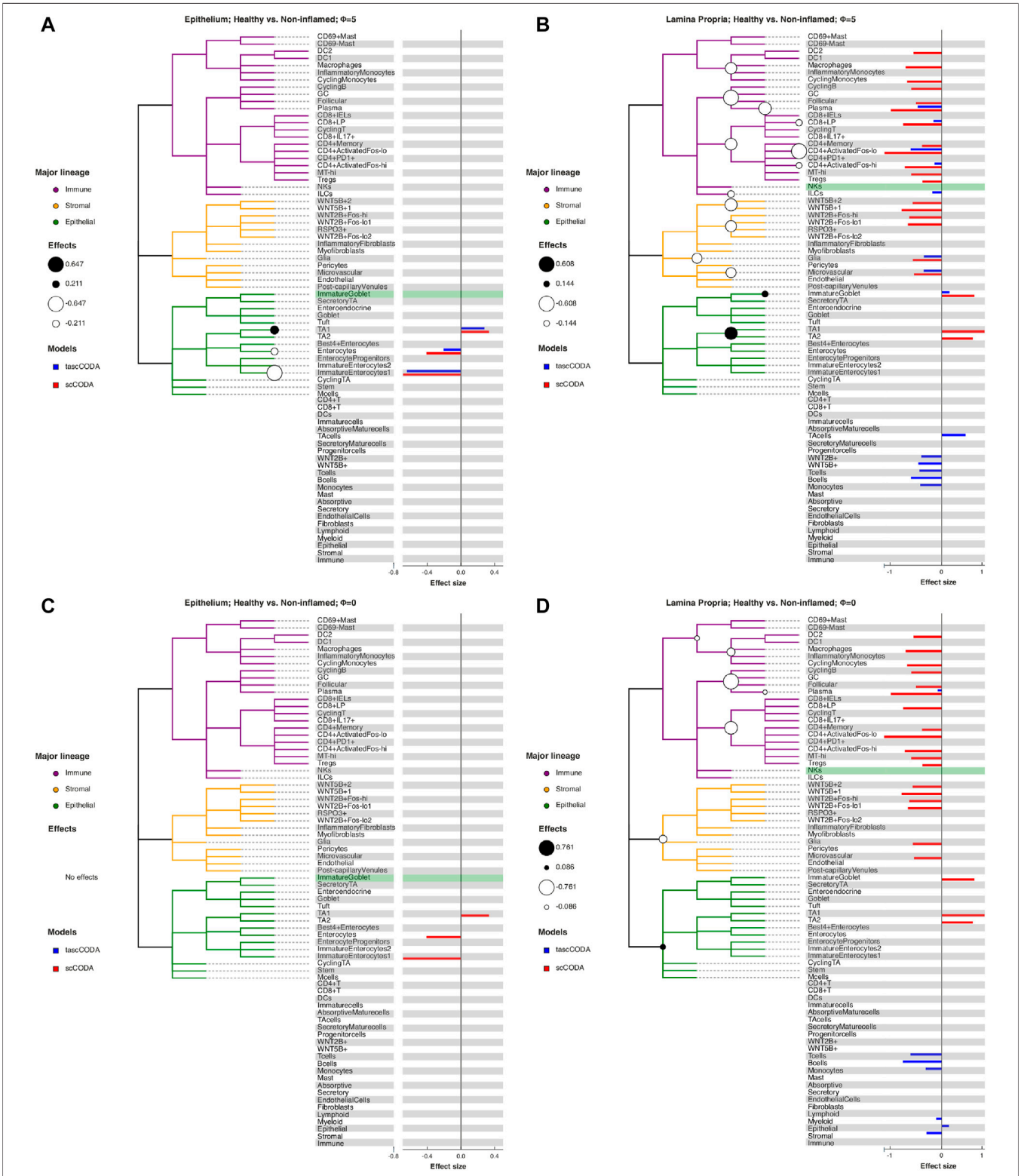


FIGURE 5 | Behavior of tascCODA on scRNA-seq data for different values of ϕ. All plots show the comparison of healthy control samples to non-inflamed tissue samples of UC patients in the data from Smillie et al. (2019). White and black circles on the cell lineage tree show the effects found by tascCODA, which are also shown as blue bars on the right side of each plot. The bars below the tree depict effects on internal nodes, with lower positions in the diagram corresponding to nodes closer to the root. For comparison, the red bars indicate effects found by scCODA, which only operates on the tips of the tree. The green-shaded area shows the reference cell type that was used for both models. **(A)** When $\phi = 5$, tascCODA prefers placing effects near the tips of the tree and finds the exact same solution as scCODA for the Epithelium data. **(B)** In the Lamina Propria, tascCODA places some effects on internal nodes, resulting in a sparser solution than the one obtained by scCODA (14 vs. 21 credible effects). **(C)** When $\phi = 0$, tascCODA finds no credible effects in samples from the Epithelium, and **(D)** only seven effects are necessary to summarize the large number of effects found by scCODA when looking at samples from the Lamina Propria.

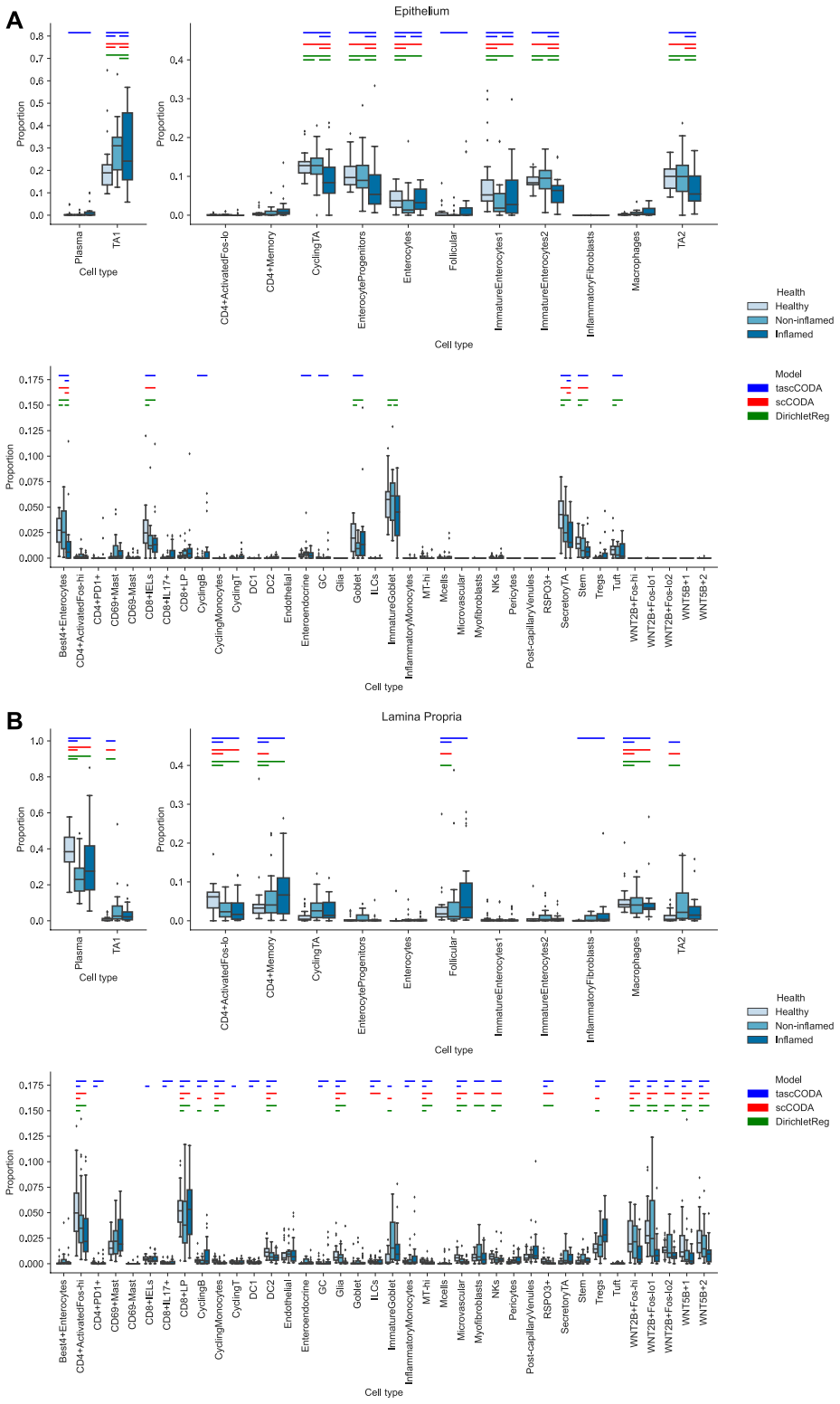


FIGURE 6 | Comparison of differentially abundant cell types found by tascCODA (blue, $\phi = 5$), scCODA (red, FDR = 0.05), and Dirichlet regression (green, adjusted $p_{adj} < 0.05$) between biopsies of healthy, non-inflamed and inflamed tissue. Colored bars for each method indicate that a credible change was found. **(A)** Among samples from the intestinal epithelium, tascCODA and Dirichlet regression detect effects on lowly abundant epithelial cell types (Tuft, Goblet, Enteroendocrine) that were not detected by scCODA. **(B)** In the Lamina Propria, only tascCODA detects a number of effects on some of the T and B cell types.

TABLE 1 | Mean squared logarithmic error (MSLE) and number of selected effects over five cross-validation splits for tascCODA with different parametrizations ϕ and scCODA. Abbreviations for scenarios: Healthy (H), Non-inflamed (N), and Inflamed (I). With increasing ϕ, tascCODA selects more effects and on average improves its predictive power. At $\phi = 5$, tascCODA has equal or lower MSLE than scCODA and a similar number of selected effects.

	Model	tascCODA					scCODA
Scenario	ϕ	−5	−1	0	1	5	-
Epithelium - H vs. N	MSLE	142.22	142.16	142.18	138.56	134.36	134.96
	Effects	0.0	0.0	0.0	1.2	3.2	2.4
Epithelium - H vs. I	MSLE	167.46	163.60	160.68	158.06	154.64	154.44
	Effects	0.0	1.6	2.6	3.2	8.2	10.8
Epithelium - N vs. I	MSLE	173.94	174.10	174.10	175.86	177.26	174.78
	Effects	0.0	0.0	0.0	0.2	3.6	5.2
LP - H vs. N	MSLE	162.76	157.62	155.16	152.80	149.58	154.02
	Effects	0.4	1.8	3.0	6.2	16.0	14.4
LP - H vs. I	MSLE	188.58	182.96	178.88	176.02	173.32	173.40
	Effects	0.0	1.8	4.8	7.8	17.8	17.4
LP - N vs. I	MSLE	219.72	219.70	219.66	219.68	216.76	218.62
	Effects	0.0	0.0	0.0	0.0	1.4	0.4

(**Figure 5D**). Again, B and T cells were found as the cell lineages that undergo the largest change between healthy and non-inflamed UC biopsies. When testing healthy versus inflamed, and non-inflamed versus inflamed biopsies, tascCODA also detected more detailed results when $\phi = 5$, and found fewer, more generalizing effects with $\phi = 0$ (**Supplementary Figures S11, S12**; **Supplementary Tables S1–S3**).

3.2.2 Analysis of the Human Gut Microbiome Under Irritable Bowel Syndrome

We next considered a microbiome data example and focused on another chronic disorder of the human gut, the Irritable Bowel Syndrome (IBS). IBS is a functional bowel disorder characterized by frequent abdominal pain, alteration of stool morphology and/or frequency, with the absence of other gastrointestinal diseases (i.e. colorectal cancer, inflammatory bowel disease). It is estimated that about 10% of the general population experience symptoms that can be classified as a subtype of Irritable Bowel Syndrome, which include IBS-C (constipation), IBS-D (diarrhea), IBS-M (mixed), or unspecified IBS (Ford et al. (2017)). While the exact sources of the disease can be manifold, it has been hypothesized that the gastroenterological symptoms may be caused by a disturbed composition of the gut microbiome (Duan et al. (2019); Ford et al. (2017)).

In particular, we analyzed 16S rRNA sequencing data of stool samples collected from IBS patients and healthy controls, which were obtained by Labus et al. (2017). The dataset consists of $n = 52$ samples, with 23 healthy controls, and 29 IBS patients separated into 11 subjects with constipation (IBS-C), 10 subjects with diarrhea (IBS-D), 6 subjects with mixed symptoms (IBS-M), and 2 subjects with unspecified symptoms. Further, metadata information about age, sex and BMI of most subjects is available. We re-processed the raw 16S rRNA sequences with DADA2, version 1.21.0 (Callahan et al. (2016)) and did taxonomic assignment via the Silva database, version 138.1 (Quast et al. (2013); Yilmaz et al. (2014)), yielding a final count table with 709 ASVs along with a taxonomic tree (**Supplementary Data 1.3.2**). This data was then aggregated at the genus level, resulting in a total of $p = 91$ known genera.

We applied tascCODA to the genus-level data, comparing healthy and IBS subjects. To showcase the flexibility of tascCODA, we analyzed the data with different covariate setups, by including the other available metadata variables. As a reference genus for scCODA and tascCODA, we chose *Alistipes*, since it is a genus with relatively high presence and rather low dispersion. For all analyses on this dataset, we decreased the mean shrinkage in tascCODA to $\lambda_1 = 1$, allowing us to find more subtle effects.

We first used tascCODA to analyze the differences in the gut microbial composition between healthy controls and IBS patients (**Figure 7**, **Supplementary Table S4**). Favoring generalization with $\phi = -5$, we found only a small decrease of the phylum Firmicutes (**Figure 7A**). In the unbiased setting ($\phi = 0$), the previous effect on the phylum level was substantiated to the Oscillospirales order. Additionally, decreases of the *Parabacteroides* and *Bacteroides* genera are found (**Figure 7B**). Setting $\phi = 5$, thus favoring detailed results, we discovered a decrease of the Ruminococcaceae family, a subgroup of Oscillospirales, and multiple decreasing genera with the strongest effects on *Parabacteroides* and *Bacteroides* (**Figure 7C**). For comparison, we also applied scCODA (FDR = 0.1) to the same dataset, which also discovered a decrease of *Parabacteroides* and *Bacteroides*, as well as three genera in the Ruminococcaceae family. A decrease of *Parabacteroides* in a subset of IBS patients was also found by Labus et al. (2017). Also, a relative decrease of the order Bacteroidales, which includes *Parabacteroides* and *Bacteroides*, was reported by Nagel et al. (2016) and Jeffery et al. (2012). Decreasing shares of Ruminococcaceae were also connected to IBS in multiple studies (Durbán et al., 2012; Pozuelo et al., 2015).

To highlight the flexibilty of tascCODA, we next tried to discover changes in the gut microbiome related to age, BMI, gender, and IBS subtype. Before applying tascCODA, we min-max normalized the two former covariates to obtain a common scale for all covariates. We excluded three samples with missing information on BMI. We conducted every analysis three times with $\phi = -5, 0, 5$. When testing for changes related to one of age, gender, or BMI alone, tascCODA

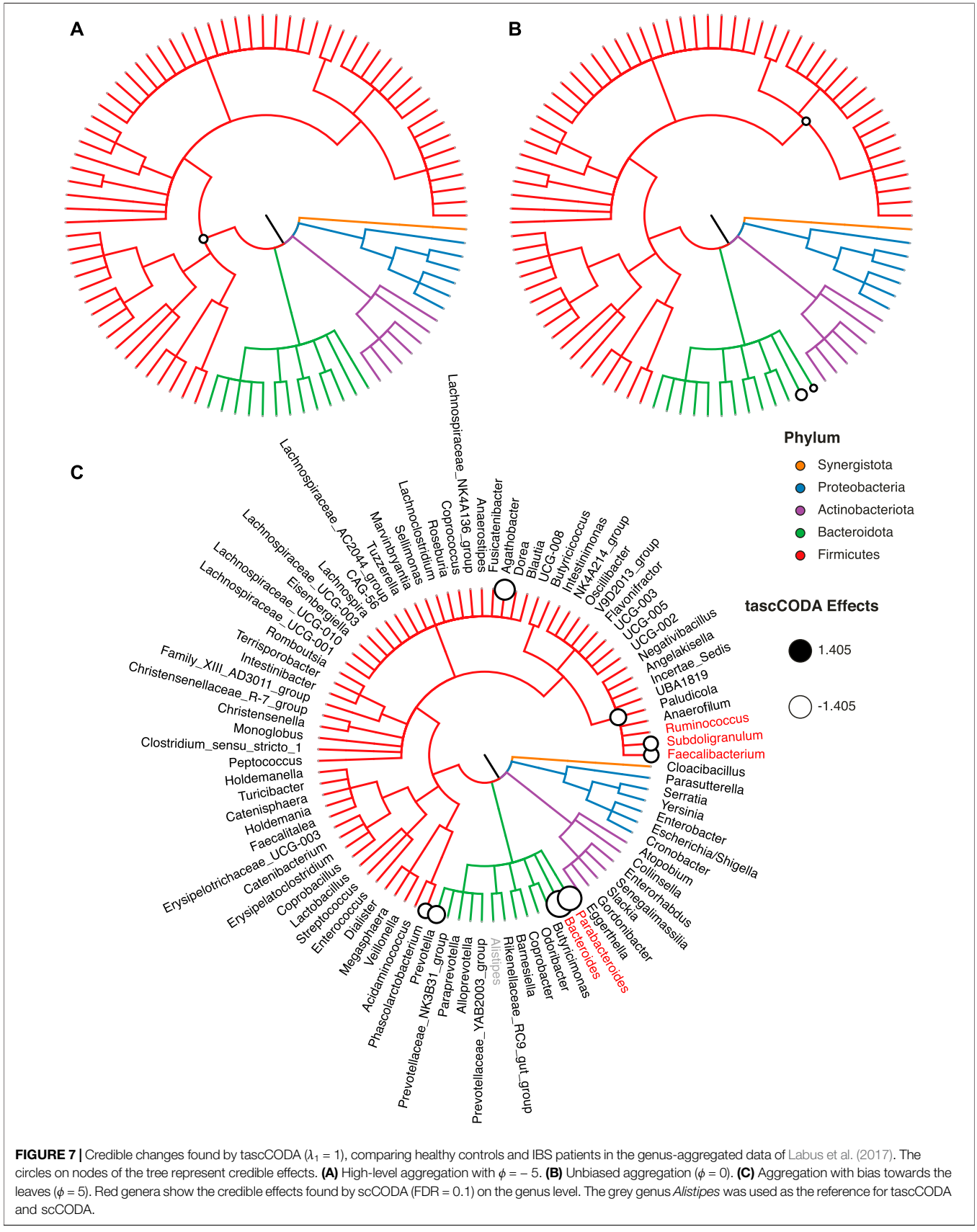


FIGURE 7 | Credible changes found by tascCODA ($\lambda_1 = 1$), comparing healthy controls and IBS patients in the genus-aggregated data of Labus et al. (2017). The circles on nodes of the tree represent credible effects. **(A)** High-level aggregation with $\phi = -5$. **(B)** Unbiased aggregation ($\phi = 0$). **(C)** Aggregation with bias towards the leaves ($\phi = 5$). Red genera show the credible effects found by scCODA (FDR = 0.1) on the genus level. The grey genus *Alistipes* was used as the reference for tascCODA and scCODA.

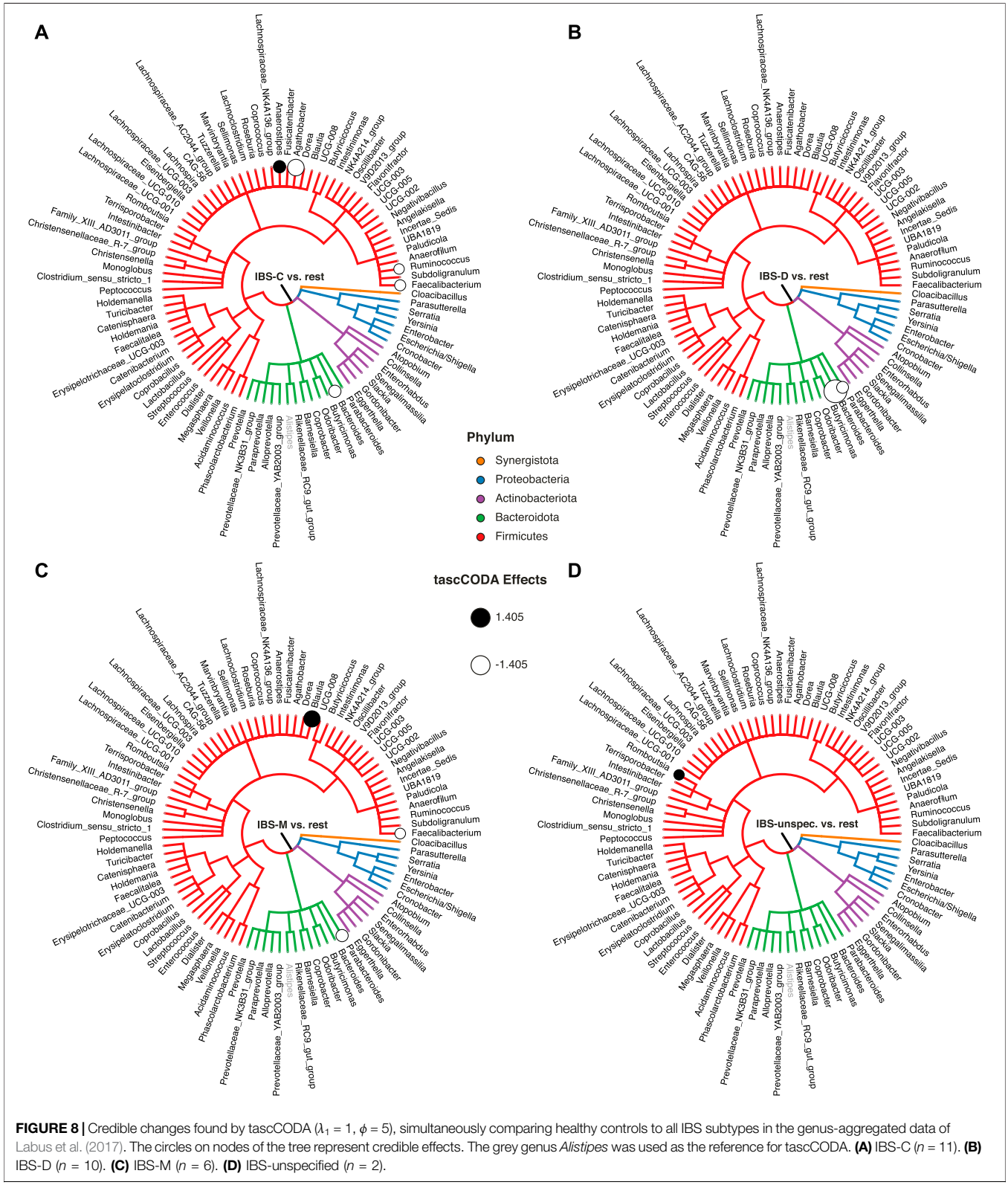


FIGURE 8 | Credible changes found by tascCODA ($\lambda_1 = 1$, $\phi = 5$), simultaneously comparing healthy controls to all IBS subtypes in the genus-aggregated data of Labus et al. (2017). The circles on nodes of the tree represent credible effects. The grey genus *Alistipes* was used as the reference for tascCODA. **(A)** IBS-C ($n = 11$). **(B)** IBS-D ($n = 10$). **(C)** IBS-M ($n = 6$). **(D)** IBS-unspecified ($n = 2$).

was not able to discover any credible differences for any aggregation bias. When testing on all four covariates together, excluding interactions, tascCODA only reported credible changes in the microbiome with respect to the IBS subtype. Finally, including all possible variables, interactions revealed that while a general negative effect was found independent of gender, male IBS-D patients had a larger depletion of *Bacteroides* than female patients.

Next, we restricted our analysis to testing for changes between the four IBS subtypes and all other samples. The results shown in **Figure 8** and **Supplementary Table S5** were obtained with $\phi = 5$. For patients experiencing constipation (IBS-C, **Figure 8A**), decreases of *Agathobacter*, *Bacteroides*, *Ruminococcus*, and *Faecalibacterium*, as well as an increase of *Anaerostipes* were found by tascCODA. Conversely, diarrhea (IBS-D, **Figure 8B**) was associated with a decrease in *Parabacteroides*, as well as a large decrease in *Bacteroides*. Patients with mixed symptoms (IBS-M, **Figure 8C**) were found to have increased numbers of *Blautia*, in addition to a decrease of *Parabacteroides* and *Faecalibacterium*, which each match with the observations related to one of the two previous conditions. Finally, only a small increase of *Romboutsia* was associated to IBS with unspecified symptoms (IBS-unspecified, **Figure 8D**).

4 DISCUSSION

Associating changes in the structure of microbial communities or cell type compositions with host or environmental covariates are commonly investigated with amplicon or single-cell RNA sequencing. With tascCODA, we have presented a fully Bayesian method to determine such compositional changes that acknowledges the hierarchical structure of the underlying microbial or cell type abundances and simultaneously accounts for the compositional nature of the data. By introducing tree-based penalization that adapts to the structure of the tree, the tascCODA model is able to accurately identify group-level changes with fewer parameters than traditional individual feature-based approaches. Thanks to a scaled variant of the spike-and-slab lasso prior (Ročková and George (2018)), we were able to obtain sparse solutions that can favor high-level aggregations or more detailed effects on a dynamic range characterized by a single scaling parameter ϕ. The tascCODA Python package seamlessly integrates into the *scanpy* environment for scRNA-seq (Wolf et al. (2018)) and allows Bayesian regression-like analyses with flexible covariate structures.

Through its ability to favor general trends or more detailed solutions, tascCODA is able to provide a trade-off between model sparsity and accuracy, which can be adjusted to reveal credible associations on different levels of the hierarchy. We recapitulated this behavior in synthetic benchmark scenarios, where focusing on low aggregation levels allowed tascCODA to outperform state-of-the-art methods in a differential abundance testing setup, while effects that influenced the majority of features were recovered with greater accuracy when we favored generalizing solutions. The aggregation property further allows for more interpretable models, detecting group-specific changes in the cell lineage or microbial taxonomy. For instance, tascCODA determined B and T cells as the main factors in cell composition changes of the Lamina Propria of Ulcerative Colitis patients, while inflamed epithelial tissue biopsies showed a depletion of Enterocytes.

Second, tascCODA can accommodate any linear combination of normalized covariates, allowing for multi-faceted analysis of complex relationships, while still producing highly sparse and interpretable solutions. On synthetic data, we showed that tascCODA was able to accurately distinguish the influence of two covariates that perturbed the data in different ways. While we did not detect credible relationships with the covariates age, sex and BMI, tascCODA was also able to simultaneously identify characteristic shifts in the gut microbiome for each subtype of Irritable Bowel Syndrome.

The application range of tascCODA extends beyond the taxonomic or expert-derived cell lineage tree structures used in our real data applications. Genetically driven orderings such as phylogenetic trees or cell type hierarchies obtained from clustering algorithms, or approaches aimed at optimizing the predictiveness of the hierarchical grouping (Quinn and Erb, 2019) may provide more accurate results in differential abundance testing (see, e.g., Bichat et al. (2020) for further information).

While tascCODA provides a hierarchically adaptive extension of a classical compositional modeling framework based on a fixed aggregation level, extensions of the method could increase the application range of tascCODA. First, tascCODA does not account for the zero-inflation and overdispersion that is common in microbial abundance data on the OTU/ASV level. We avoided this challenge here by aggregating the amplicon data to the genus level. Accounting for these properties within the model, for example by using a zero-inflated Dirichlet-Multinomial model (Tang and Chen (2019)), the Tweedie family of distributions (Mallick et al. (2021)), or hard thresholding on latent weights (Ren et al. (2020)), would allow for even more fine-grained analyses. Second, the tascCODA model currently places a sparsity-inducing spike-and-slab lasso prior on all included covariates. A natural next step would be to consider some covariates as confounding variables similar to Zhou H. et al. (2021), reducing the number of latent parameters, while restricting results to a few core influence factors. Third, extending known efficient computational methods for inference of spike-and-slab lasso priors (Bai et al. (2020b); Ročková and George (2018)) to be used with our compositional modeling framework could greatly reduce the computational resources required for running tascCODA.

We believe that tascCODA, together with its implementation in Python, represents a valuable addition to the growing toolbox of compositional data modeling tools by providing a unifying statistical way to model and analyze microbial and cell population data in the presence of hierarchical side information.

DATA AVAILABILITY STATEMENT

The model is available as a Python package on github[4]. The datasets used in this study are publicly available on Single Cell Portal (accession ID SCP259) and the Short Read Archive (accession number PRJNA373876). The scripts used for data analysis and benchmark data generation can be found in the tascCODA reproducibility repository[5]. Supplemental data can be downloaded from zenodo[6].

[4]https://github.com/bio-datascience/tascCODA.
[5]https://github.com/bio-datascience/tascCODA_reproducibility.
[6]10.5281/zenodo.5302135.

AUTHOR CONTRIBUTIONS

JO developed tascCODA and conducted the simulation studies and real data analysis. SC processed the 16S rRNA sequencing data and provided biological context. CM supervised the work. JO and CM conceived the statistical model, designed the simulation and out-of-sample prediction studies and wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

CM acknowledges core funding from the Institute of Computational Biology, Helmholtz Zentrum München.

ACKNOWLEDGMENTS

We thank Maren Büttner for providing the initial processing steps in the scRNA-seq data analysis. Furthermore, we thank Jennifer S. Labus for kindly sharing additional metadata information on the IBS data. We acknowledge Michael Menden's support in supervising SC during her Master's Thesis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.766405/full#supplementary-material

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, arXiv preprint arXiv:1603. 04467.

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Ser. B (Methodological)* 44, 139–160.

Büttner, M., Ostner, J., Müller, C. L., Theis, F. J., and Schubert, B. (2020). scCODA: A Bayesian Model for Compositional Single-Cell Data Analysis. *Nat. Commun.* 12, 6876.

Bai, R., Moran, G. E., Antonelli, J. L., Chen, Y., and Boland, M. R. (2020a). Spike-and-Slab Group Lassos for Grouped Regression and Sparse Generalized Additive Models. *J. Am. Stat. Assoc.*

Bai, R., Rockova, V., and George, E. I. (2020b). Spike-and-Slab Meets LASSO: A Review of the Spike-And-Slab LASSO. *arXiv [stat.ME]*.

Betancourt, M., and Girolami, M. (2015). Hamiltonian Monte Carlo for Hierarchical Models. In Current Trends in Bayesian Methodology with Applications. Chapman and Hall/CRC, 79–101.

Bichat, A., Plassais, J., Ambroise, C., and Mariadassou, M. (2020). Incorporating Phylogenetic Information in Microbiome Differential Abundance Studies Has No Effect on Detection Power and FDR Control. *Front. Microbiol.* 11, 649.

Bien, J., Yan, X., Simpson, L., and Müller, C. L. (2021). Tree-aggregated Predictive Modeling of Microbiome Data. *Sci. Rep.* 11, 14505. .

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: High-Resolution Sample Inference from Illumina Amplicon Data. *Nat. Methods* 13, 581–583.

Chen, J., and Li, H. (2013). Variable Selection for Sparse Dirichlet-Multinomial Regression with an Application to Microbiome Data Analysis. *Ann. Appl. Stat.* 7.

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., et al. (2017). Tensorflow Distributions. *arXiv preprint*. arXiv:1711. 10604

Duan, R., Zhu, S., Wang, B., and Duan, L. (2019). Alterations of Gut Microbiota in Patients with Irritable Bowel Syndrome Based on 16S rRNA-Targeted Sequencing: A Systematic Review. *Clin. Translational Gastroenterol.* 10, e00012.

Duò, A., Robinson, M. D., and Soneson, C. (2018). A Systematic Performance Evaluation of Clustering Methods for Single-Cell Rna-Seq Data. *F1000Res* 7, 1141.

Durbán, A., Abellán, J. J., Jiménez-Hernández, N., Salgado, P., Ponce, M., Ponce, J., et al. (2012). Structural Alterations of Faecal and Mucosa-Associated Bacterial Communities in Irritable Bowel Syndrome. *Environ. Microbiol. Rep.* 4, 242–247.

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., and Gloor, G. B. (2014). Unifying the Analysis of High-Throughput Sequencing Datasets: Characterizing RNA-Seq, 16S rRNA Gene Sequencing and Selective Growth Experiments by Compositional Data Analysis. *Microbiome* 2, 15.

Ford, A. C., Lacy, B. E., and Talley, N. J. (2017). Irritable Bowel Syndrome. *N. Engl. J. Med.* 376, 2566–2578.

Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., et al. (2014). The Treatment-Naive Microbiome in New-Onset Crohn's Disease. *Cell Host & Microbe* 15, 382–392.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* 8, 2224.

Gordon-Rodriguez, E., Quinn, T. P., and Cunningham, J. P. (2021). Learning Sparse Log-Ratios for High-Throughput Sequencing Data. *bioRxiv*. doi:10.1101/2021.02.11.430695

Griffiths, J. A., Scialdone, A., and Marioni, J. C. (2018). Using Single-Cell Genomics to Understand Developmental Processes and Cell Fate Decisions. *Mol. Syst. Biol.* 14, e8046.

Hawinkel, S., Mattiello, F., Bijnens, L., and Thas, O. (2019). A Broken Promise: Microbiome Differential Abundance Methods Do Not Control the False Discovery Rate. *Brief. Bioinform.* 20, 210–221.

He, S., Wang, L. H., Liu, Y., Li, Y. Q., Chen, H. T., Xu, J. H., et al. (2020). Single-cell Transcriptome Profiling of an Adult Human Cell Atlas of 15 Major Organs. *Genome Biol.* 21, 294.

Holmén, N., Lundgren, A., Lundin, S., Bergin, A.-M., Rudin, A., Sjövall, H., et al. (2006). Functional CD4+CD25high Regulatory T Cells Are Enriched in the Colonic Mucosa of Patients with Active Ulcerative Colitis and Increase with Disease Activity. *Inflamm. Bowel Dis.* 12, 447–456.

Homan, M. D., and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* 15, 1593–1623.

Human Microbiome Project Consortium (2012). Structure, Function and Diversity of the Healthy Human Microbiome. *Nature* 486, 207–214.

Jeffery, I. B., O'Toole, P. W., Öhman, L., Claesson, M. J., Deane, J., Quigley, E. M. M., et al. (2012). An Irritable Bowel Syndrome Subtype Defined by Species-specific Alterations in Faecal Microbiota. *Gut* 61, 997–1006.

Karlsson, M., Zhang, C., Méar, L., Zhong, W., Digre, A., Katona, B., et al. (2021). A Single–Cell Type Transcriptomics Map of Human Tissues. *Sci. Adv.* 7, 2169.

Kumar, R., Carroll, C., Hartikainen, A., and Martin, O. (2019). ArviZ a Unified Library for Exploratory Analysis of Bayesian Models in python. *Joss* 4, 1143.

Labus, J. S., Hollister, E. B., Jacobs, J., Kirbach, K., Oezguen, N., Gupta, A., et al. (2017). Differences in Gut Microbial Composition Correlate with Regional Brain Volumes in Irritable Bowel Syndrome. *Microbiome* 5, 49.

Lin, H., and Peddada, S. D. (2020). Analysis of Compositions of Microbiomes with Bias Correction. *Nat. Commun.* 11, 3514.

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., et al. (2017). Strains, Functions and Dynamics in the Expanded Human Microbiome Project. *Nature* 550, 61–66.

Luecken, M. D., and Theis, F. J. (2019). Current Best Practices in Single-Cell Rna-Seq Analysis: a Tutorial. *Mol. Syst. Biol.* 15, e8746.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161, 1202–1214.

Maier, M. J. (2014). *DirichletReg: Dirichlet Regression for Compositional Data in R*. Research Report Series 125. Vienna, Austria: Vienna University of Economics and Business.

Mallick, H., Chatterjee, S., Chowdhury, S., Chatterjee, S., Rahnavard, A., and Hicks, S. C. (2021). Differential Expression of Single-Cell RNA-Seq Data Using Tweedie Models.

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., and Peddada, S. D. (2015). Analysis of Composition of Microbiomes: a Novel Method for Studying Microbial Composition. *Microb. Ecol. Health Dis.* 26, 27663.

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., et al. (2018). American Gut: an Open Platform for Citizen Science Microbiome Research. *Msystems* 3, e00031–18.

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., et al. (2012). An Improved Greengenes Taxonomy with Explicit Ranks for Ecological and Evolutionary Analyses of Bacteria and Archaea. *ISME J.* 6, 610–618.

McKinney, W. (2010). Data Structures for Statistical Computing in python. In Proceedings of the 9th Python in Science Conference. (Austin, Texas, USA: SciPy).

Nagel, R., Traub, R. J., Allcock, R. J. N., Kwan, M. M. S., and Bielefeldt-Ohmann, H. (2016). Comparison of Faecal Microbiota in Blastocystis-Positive and Blastocystis-Negative Irritable Bowel Syndrome Patients. *Microbiome* 4, 47.

Nesterov, Y. (2009). Primal-dual Subgradient Methods for Convex Problems. *Math. Program* 120, 221–259.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R Language. *Bioinformatics* 20, 289–290.

Pozuelo, M., Panda, S., Santiago, A., Mendez, S., Accarino, A., Santos, J., et al. (2015). Reduction of Butyrate- and Methane-Producing Microorganisms in Patients with Irritable Bowel Syndrome. *Sci. Rep.* 5, 12693.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools. *Nucleic Acids Res.* 41, D590–D596.

Quinn, T. P., and Erb, I. (2019). Using Balances to Engineer Features for the Classification of Health Biomarkers: a New Approach to Balance Selection. *bioRxiv*. doi:10.1101/600122

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., et al. (2017). The Human Cell Atlas. *elife* 6, e27041.

Ren, B., Bacallado, S., Favaro, S., Vatanen, T., Huttenhower, C., and Trippa, L. (2020). Bayesian Mixed Effects Models for Zero-Inflated Compositions in Microbiome Data Analysis. *Ann. Appl. Stat.* 14, 494–517.

Ročková, V., and George, E. I. (2018). The Spike-And-Slab LASSO. *J. Am. Stat. Assoc.* 113, 431–444.

Round, J. L., and Palm, N. W. (2018). Causal Effects of the Microbiota on Immune-Mediated Diseases. *Sci. Immunol.* 3.

Schliep, K. P. (2010). Phangorn: Phylogenetic Analysis in R. *Bioinformatics* 27, 592–593.

Scott, J. G., and Berger, J. O. (2010). Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem. *Ann. Statist.* 38, 2587–2619.

Sender, R., Fuchs, S., and Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *Plos Biol.* 14, e1002533–14.

Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., et al. (2013). Single-cell Transcriptomics Reveals Bimodality in Expression and Splicing in Immune Cells. *Nature* 498, 236–240.

Silverman, J. D., Washburne, A. D., Mukherjee, S., and David, L. A. (2017). A Phylogenetic Transform Enhances Analysis of Compositional Microbiota Data. *Elife* 6.

Smillie, C. S., Biton, M., Ordovas-Montanes, J., Sullivan, K. M., Burgin, G., Graham, D. B., et al. (2019). Intra- and Inter-cellular Rewiring of the Human colon during Ulcerative Colitis. *Cell* 178, 714–730.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). Mrna-Seq Whole-Transcriptome Analysis of a Single Cell. *Nat. Methods* 6, 377–382.

Tang, Z.-Z., and Chen, G. (2019). Zero-inflated Generalized Dirichlet Multinomial Regression Model for Microbiome Compositional Data Analysis. *Biostatistics* 20, 698–713.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)* 58, 267–288.

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing Well-Connected Communities. *Sci. Rep.* 9, 5233.

Trapnell, C. (2015). Defining Cell Types and States with Single-Cell Genomics. *Genome Res.* 25, 1491–1498.

Tsoucas, D., Dong, R., Chen, H., Zhu, Q., Guo, G., and Yuan, G. C. (2019). Accurate Estimation of Cell-type Composition from Gene Expression Data. *Nat. Commun.* 10, 2975–2979.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The Human Microbiome Project. *Nature* 449, 804–810.

Wadsworth, W. D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S. A., and Vannucci, M. (2017). An Integrative Bayesian Dirichlet-Multinomial Regression Model for the Analysis of Taxonomic Abundances in Microbiome Data. *BMC Bioinformatics* 18, 94.

Wang, T., and Zhao, H. (2017). A Dirichlet-Tree Multinomial Regression Model for Associating Dietary Nutrients with Gut Microorganisms. *Biom* 73, 792–801.

Wang, Z., Mao, J., and Ma, L. (2021). Logistic-tree normal Model for Microbiome Compositions. *arXiv [stat.ME]*.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis. *Genome Biol.* 19, 15.

Yan, X., and Bien, J. (2021). Rare Feature Selection in High Dimensions. *J. Am. Stat. Assoc.* 116, 887–900.

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., et al. (2014). The SILVA and "All-Species Living Tree Project (LTP)" Taxonomic Frameworks. *Nucl. Acids Res.* 42, D643–D648.

Zhou, C., Zhao, H., and Wang, T. (2021a). Transformation and Differential Abundance Analysis of Microbiome Data Incorporating Phylogeny. *Bioinformatics*.

Zhou, H., Zhang, X., He, K., and Chen, J. (2021b). LinDA: Linear Models for Differential Abundance Analysis of Microbiome Compositional Data. *arXiv [stat.ME]*.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary Material

1 SUPPLEMENTARY DATA

1.1 Notation overview

This section gives an overview over the inputs and parameters used by tascCODA:

Data inputs

- $Y \in \mathbb{R}^{n \times p}$ is the count matrix of features $j = 1 \dots p$ in samples $i = 1 \dots n$. $\bar{Y}_i = \sum_{j=1}^{p} Y_{i,j}$ is the sequencing depth of sample i.
- $X \in \mathbb{R}^{n \times d}$ is the covariate matrix of covariates $l = 1 \dots d$ for samples $i = 1 \dots n$.
- $\mathcal{T}$ is a multifurcating tree structure with p leaves and t internal nodes defined by the ancestor matrix $A \in \{0, 1\}^{p \times v}$, with $v = p + t$

Latent parameters

- $\mathbf{a}_i = (a_{1,i}, \dots a_{p,i}); a_{j,i} \geq 0$ is the probability vector of the Dirichlet-Multinomial distribution for sample i.
- α_j is the base (intercept) parameter for feature j.
- $\beta_{l,j}$ is the effect of covariate l on feature j.
- $\hat{\beta}_{l,k}$ is the effect of covariate l on tree node k.
- $\tilde{\beta}_{0,l,k}$ is the spike portion of the spike-and-slab LASSO prior for covariate l and tree node k with parameters $\sigma_{0,l,k}$ and $b_{0,l,k}$.
- $\tilde{\beta}_{1,l,k}$ is the slab portion of the spike-and-slab LASSO prior for covariate l and tree node $k$with parameters $\sigma_{1,l,k}$ and $b_{1,l,k}$.
- θ is the mixture coefficient of the spike-and-slab LASSO prior.

Tuning parameters/hyperparameters

- λ_0 is the shrinkage parameter for the spike portion, default $\lambda_0 = 50$.
- $\lambda_{1,k}$ is the node-specific shrinkage parameter for the slab portion of the prior on node k, with mean value λ_1, default $\lambda_1 = 5$.
- ϕ is the aggregation bias parameter for scaling the slab shrinkage $\lambda_{1,k}$

1.2 Hyperparameters for the spike-and-slab LASSO prior

We want to shed some additional light on the role of the hyperparameters $\lambda_0, \lambda_1, \theta$ in the spike-and-slab LASSO prior (Ročková and George (2018)). For simplicity and because the model is symmetric with respect to the covariates, we assume $d = 1$ and thus refrain from indexing parameters with the covariate. For one node $\hat{\beta}_k$, the prior is a mixture of two double-exponential distributions $\psi_0(\hat{\beta}_k)$ and $\psi_1(\hat{\beta}_k)$ (Figure S1A) whose share is determined by θ:

$$p(\hat{\beta}_k|\theta) = \theta\psi_1(\hat{\beta}_k) + (1-\theta)\psi_0(\hat{\beta}_k) \tag{S1}$$

$$\psi_1(\hat{\beta}_k) = \frac{\lambda_1}{2}e^{-\lambda_1|\hat{\beta}_k|} \tag{S2}$$

$$\psi_0(\hat{\beta}_k) = \frac{\lambda_0}{2}e^{-\lambda_0|\hat{\beta}_k|} \tag{S3}$$

The double-exponential density (S2) has a peak at zero for large values of λ, which decreases with λ (Bai et al. (2020)). Thus, setting $\lambda_0 \gg \lambda_1$ in the mixture density (S1) results in a product of a peaked "spike" (ψ_0) and a diffuse "slab" (ψ_1) component (Figure S1B). Interestingly, Ročková and George (2018) showed that the spike-and-slab LASSO prior can be reformulated as a penalized likelihood method that is, for fixed θ:

$$pen(\hat{\beta}_k|\theta) = -\lambda_1|\hat{\beta}_k| + \log(\frac{p^*_\theta(0)}{p^*_\theta(\hat{\beta}_k)}) \tag{S4}$$

where

$$p^*_\theta(b) = \frac{\theta\frac{\lambda_1}{2}e^{-\lambda_1|b|}}{\theta\frac{\lambda_1}{2}e^{-\lambda_1|b|} + (1-\theta)\frac{\lambda_0}{2}e^{-\lambda_0|b|}} \tag{S5}$$

In the case of $\lambda_0 = \lambda_1$, the log-term in (S4) vanishes, and the penalty is equivalent to the standard LASSO (Tibshirani (1996)).

After making the weight θ data-adaptive by a Beta prior (Equation (9)), we turn our attention to the double-exponential parameters. We show the influence of each parameter on the solution by simulations on one of the randomly generated datasets from the simulation study with $p = 10$ features. From Figure S2, we can see that the ground truth assumption are effects on nodes 0, 4, and 12, with the latter node affecting features 7 and 8. We first fix $\lambda_1 = 1$, and vary λ_0 on a scale between 1 and 1000. Figure S1C shows that the effects $\hat{\beta}$ quickly stabilize with the three true effects being clearly separated from all other effects, which are close to zero. This stabilization was also explained by Ročková and George (2018) and is rooted in the fact that larger values of λ_0 only narrow the spike, which does not affect the solution after some point. We can thus simply set λ_0 to a relatively large value, the default in tascCODA is $\lambda_0 = 50$. When $\lambda_0 = \lambda_1$, we can see the typical parameter curve of a LASSO model, where the true effects are the last to approach zero (Figure S1D).

Because $\lambda_1 \to \lambda_0$ approaches the $\mathcal{L}_1$ penalty of the LASSO, which will eventually force all effects towards zero, leaving $\lambda_0 = 50$ and increasing λ_1 shows a similar behavior (Figure S1E). Only the true effects are significantly larger than zero once λ_1 reaches a value of approximately 0.1. After a certain point ($\lambda_1 \approx 10$), the penalty becomes so large that all effects vanish. We utilize the regularizing behavior by scaling λ_1 depending on the number of leaves that a node influences to put a preference on nodes on different levels of the tree (Equation 10). The direction and steepness of the preference is expressed by the parameter ϕ, with $\phi = 0$ giving equal treatment to all nodes. The default overall size of the penalty, $\lambda_1 = 5$, is chosen in a way that the parameters $\lambda_{1,k} \in (0, 10]$ stay in the range of values that were recommended by Ročková and George (2018) for all k. Figure S1F shows how the results change with different values of ϕ.

For $\phi <= 0$, favoring high-level aggregations, the model selects the three ground truth nodes. When $\phi > 1$, tips are penalized considerably less than internal nodes and the effect on node 12 is replaced by equal-sized effects on its children, nodes 7 and 8. Also, for $\phi < 0$, effects on nodes that are high in the tree (large k) are different from zero, but smaller than the significance threshold (dashed line), while for $\phi > 0$, this is the case for leaf nodes.

1.3 Experimental data preprocessing

1.3.1 Single-cell RNA-seq analysis of ulcerative colitis in humans

We obtained the data on ulcerative colitis from from Single Cell Portal (accession ID SCP259) and the analysis code from github. In total, the data consists of 365,492 transcriptomes from 12 healthy donors and 18 donors with UC providing non-inflamed and inflamed tissue samples. We used the 51 different cell types found in the original analysis, but considered every replicate as an independent sample, as done in a re-analysis by Büttner et al. (2020) on the same dataset. Biopsies from two different tissue regions, the Epithelium ('Epi' - 24 healthy, 21 non-inflamed, 16 inflamed samples) and the underlying Lamina Propria ('LP' - 24 healthy, non-inflamed, and inflamed samples each), were divided by enzymatic digestion. We inferred the cell lineage tree from the Methods section of Smillie et al. (2019) (Figure S11).

1.3.2 Analysis of the human gut microbiome under Irritable Bowel Syndrome

The raw 16S rRNA sequences (avilable at the Short Read Archive, accession number PRJNA373876) were re-processed using DADA2, version 1.21.0 (Callahan et al. (2016)). After primer and quality filtering (minimum read length: 150bp, maximum errors per read: 3, reads trimmed at first base with quality below: 10), inference of ASVs and removal of chimeras, the taxonomy of the inferred ASVs was determined with the Silva database, version 138.1 (Quast et al. (2013); Yilmaz et al. (2014)). Samples with a total read count of less than 500 (n=0) were discarded and ASVs assigned to Eukaryota (n=0) or belonging to an unknown Phylum (n=1) were removed, yielding a final count table with 709 ASVs along with a taxonomic tree.

2 SUPPLEMENTARY TABLES AND FIGURES

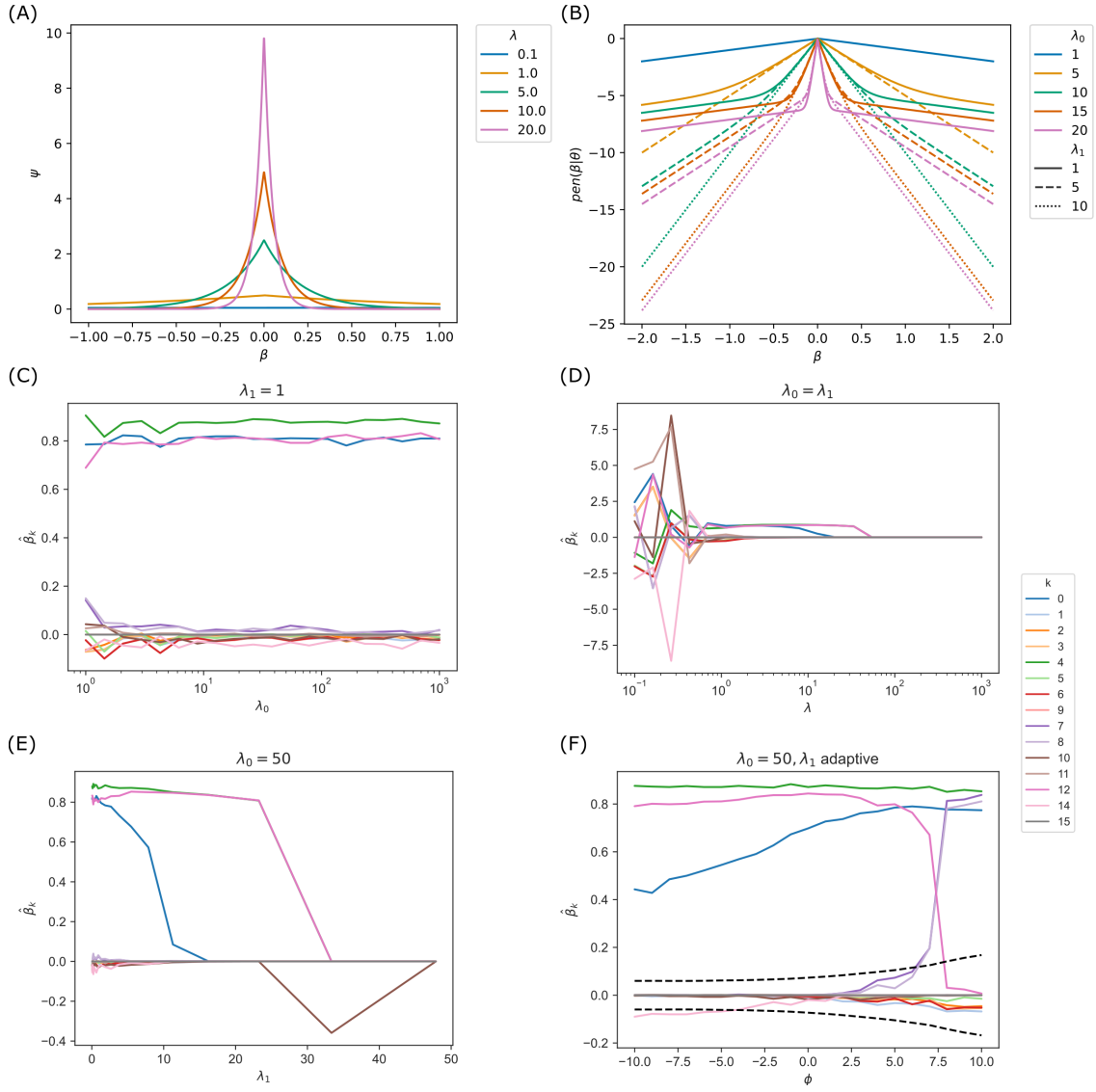


Figure S1. Parameters in the spike-and-slab LASSO penalty. **(A)** The double exponential density $\psi(\beta, \lambda)$ for different values of λ. The density becomes more peaked with increasing λ. **(B)** The likelihood penalty (Equation (S4)) introduced by different parametrizations of the spike-and-slab LASSO prior ($\theta = 0.1$). For larger effect sizes β, the penalty is driven by the slab parameter λ_1 (lines with the same style are close together). For smaller effect sizes β, the penalty is driven by the spike parameter λ_0 (lines with the same color are close together). If $\lambda_0 = \lambda_1$, the penalty is linear and equivalent to the LASSO penalty $\lambda_0\beta$. **(C-F)** Effect of different parameters on the effects $\hat{\beta}_k$ determined by tascCODA. For all simulations, a realization of the dataset in Supplementary Figure S2 was used. The nodes 13, 16 and 17 are singularities and were thus deleted before model application. **(C)** Solutions found by tascCODA when varying values of λ_0 and constant $\lambda_1 = 1$. The effects $\hat{\beta}_k$ stabilize and increasing λ_0 has no effect. **(D)** Solutions found by tascCODA in a LASSO-equivalent setting when varying values of $\lambda_0 = \lambda_1 = \lambda$. With increasing λ, more effects $\hat{\beta}_k$ go to 0. **(E)** Solutions found by tascCODA when varying values of λ_1 and constant $\lambda_0 = 50$. With increasing λ_1, a similar effect to the LASSO can be seen, where all effects are eventually approaching 0. **(F)** Solutions found by tascCODA when varying the tree level bias ϕ. $\lambda_0 = 50$, $\lambda_{1,k}$ as in Equation 10. The dashed black lines show the significance threshold (Equation 11).

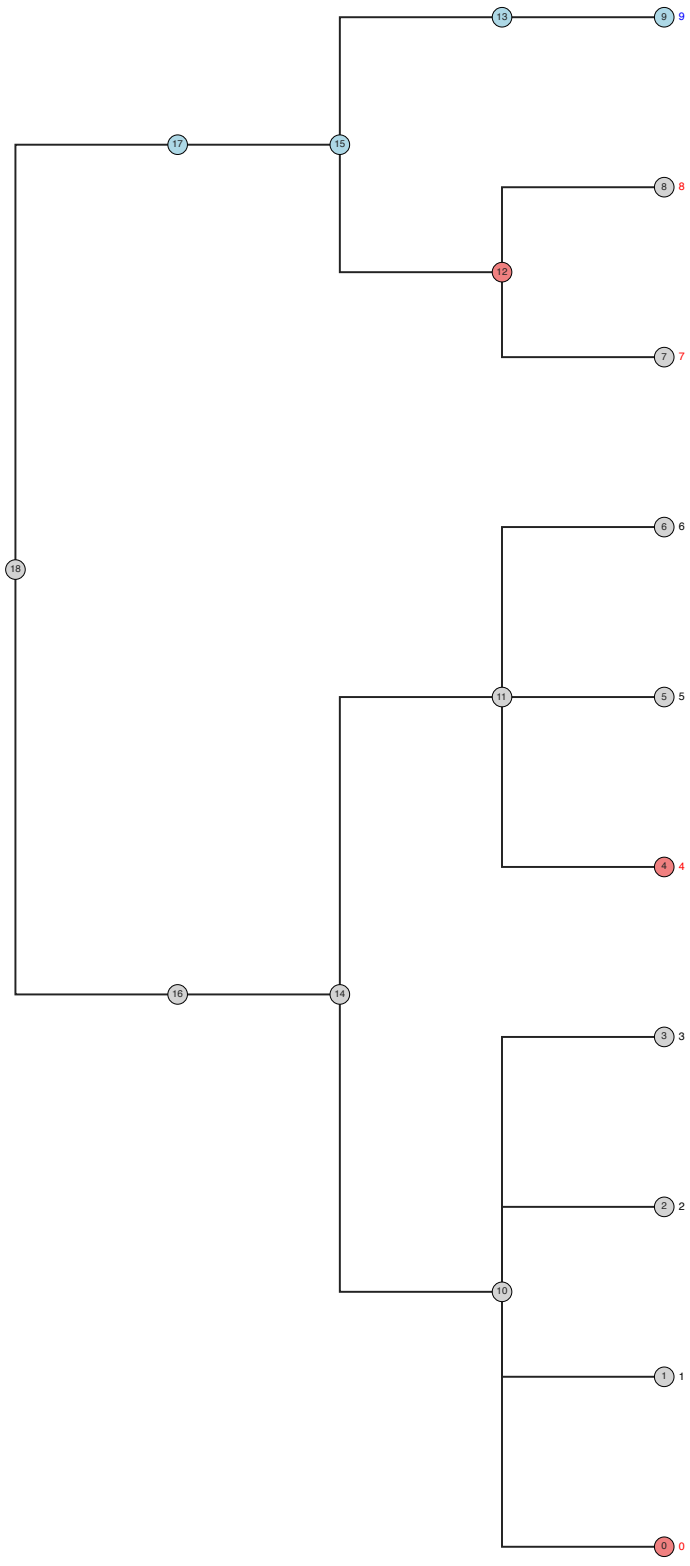


Figure S2. Randomly generated tree structure for synthetic data benchmark, $p = 10$ tips. The red nodes were selected to be affected by the condition, causing the red tips to be differentially abundant. The blue tip is the reference feature, which forces the effects on all blue nodes to be 0.

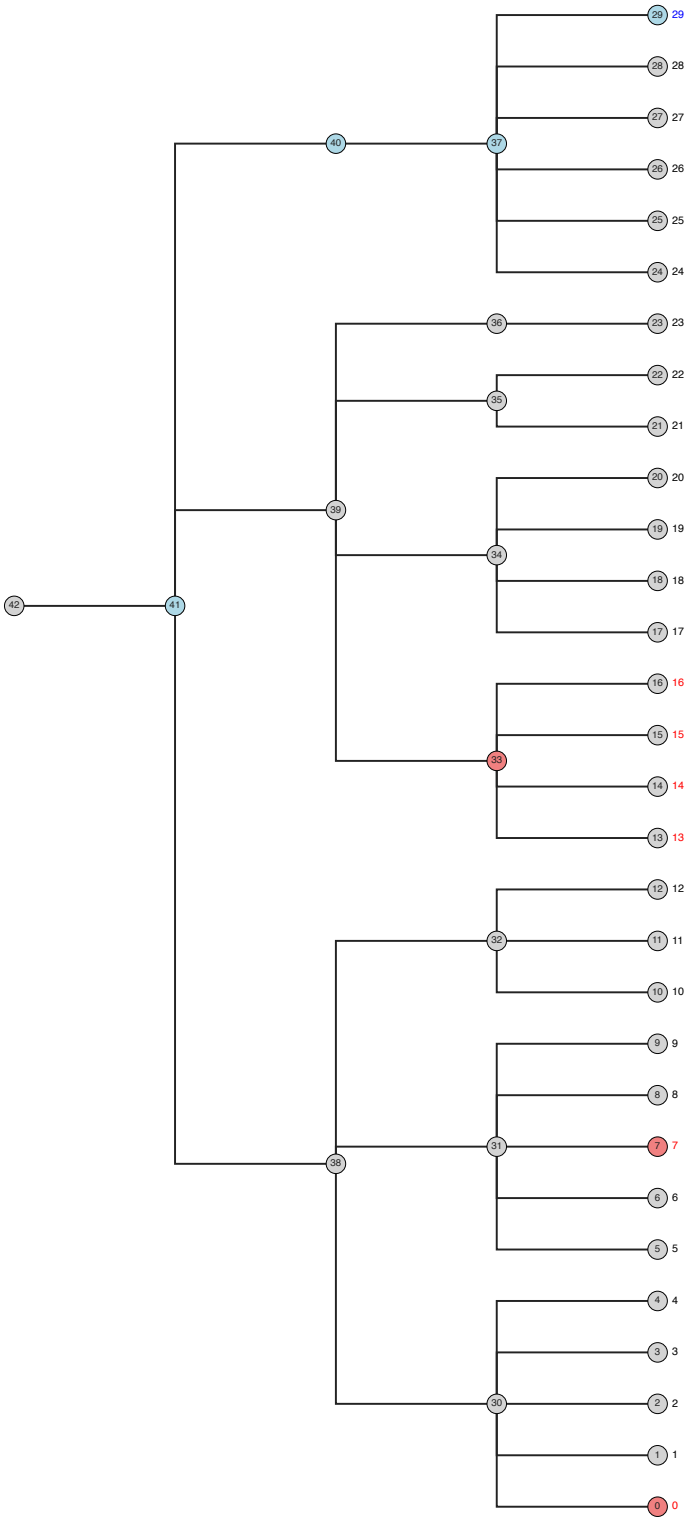


Figure S3. Randomly generated tree structure for synthetic data benchmark, $p = 30$ tips. The red nodes were selected to be affected by the condition, causing the red tips to be differentially abundant. The blue tip is the reference feature, which forces the effects on all blue nodes to be 0.

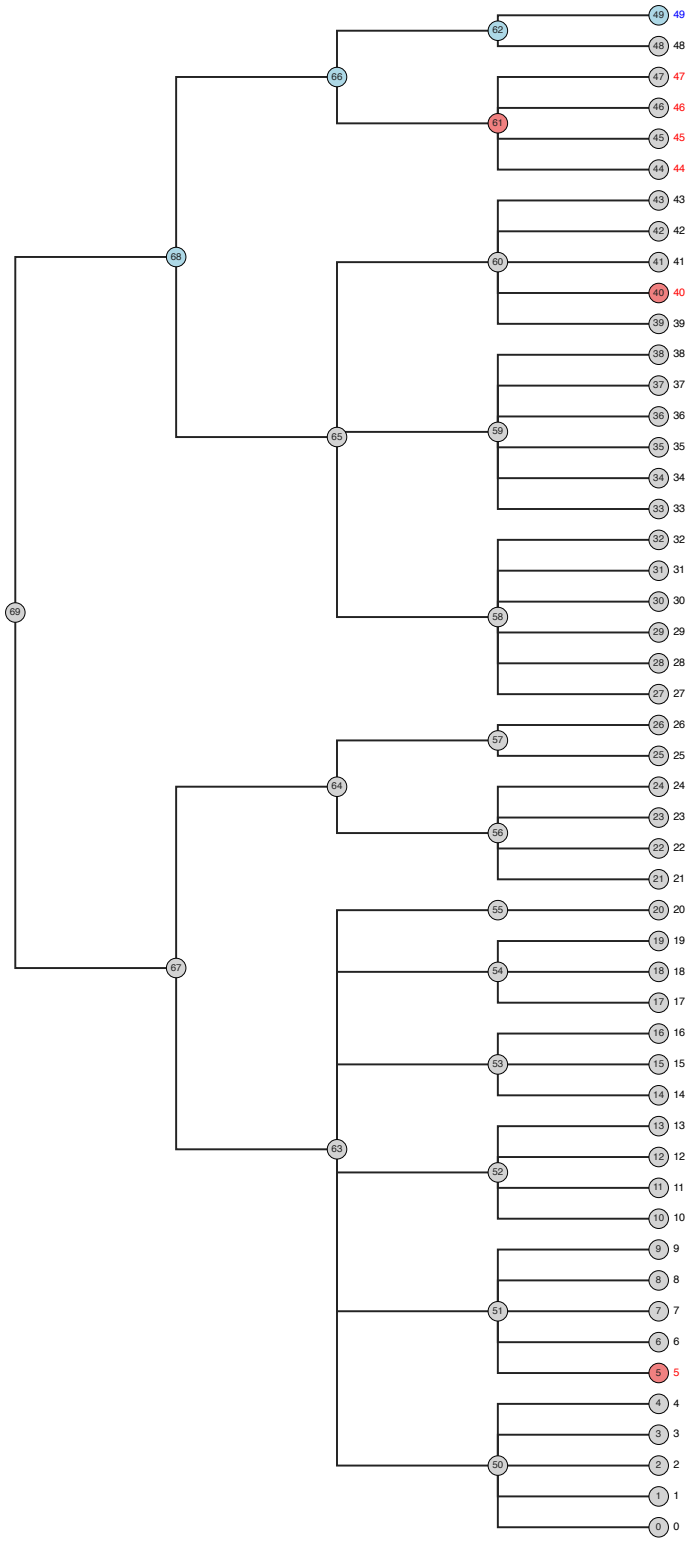


Figure S4. Randomly generated tree structure for synthetic data benchmark, $p = 50$ tips. The red nodes were selected to be affected by the condition, causing the red tips to be differentially abundant. The blue tip is the reference feature, which forces the effects on all blue nodes to be 0.

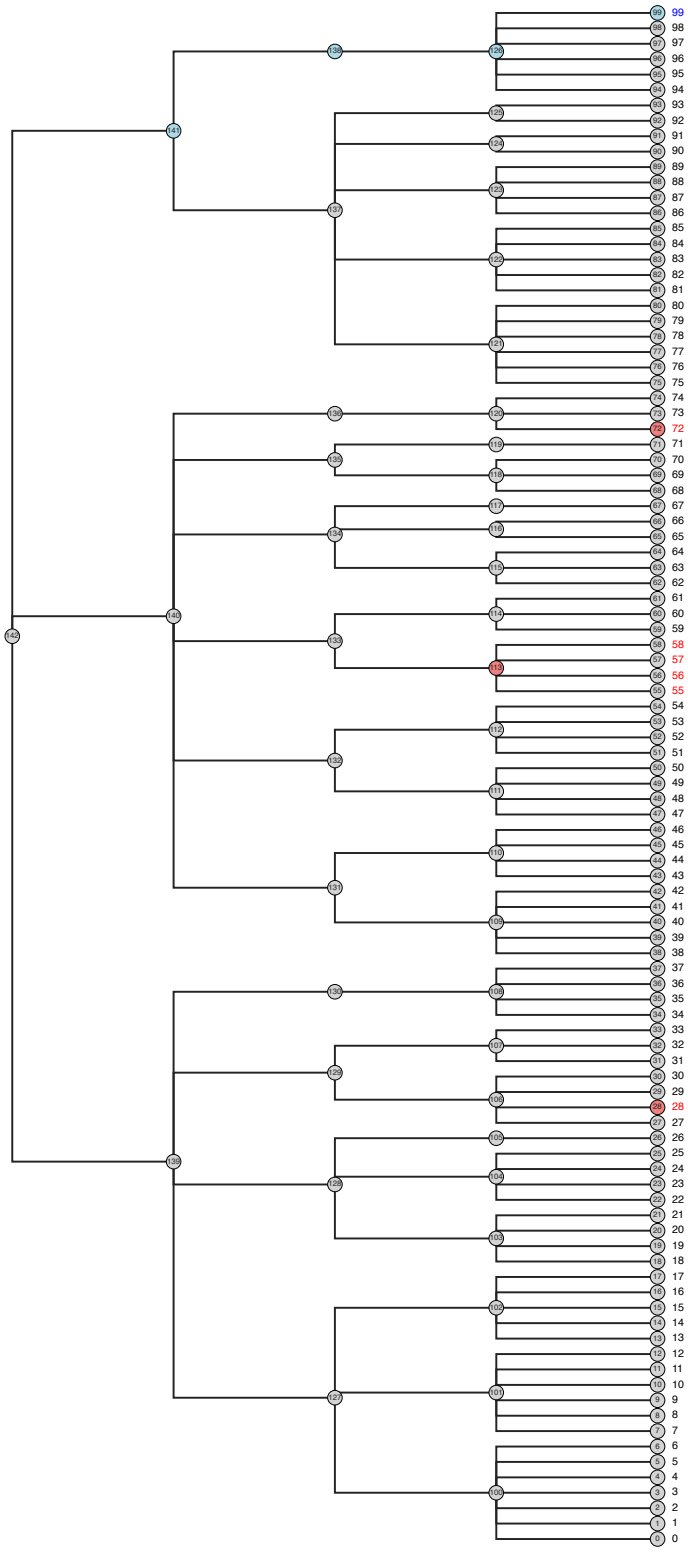


Figure S5. Randomly generated tree structure for synthetic data benchmark, $p = 100$ tips. The red nodes were selected to be affected by the condition, causing the red tips to be differentially abundant. The blue tip is the reference feature, which forces the effects on all blue nodes to be 0.

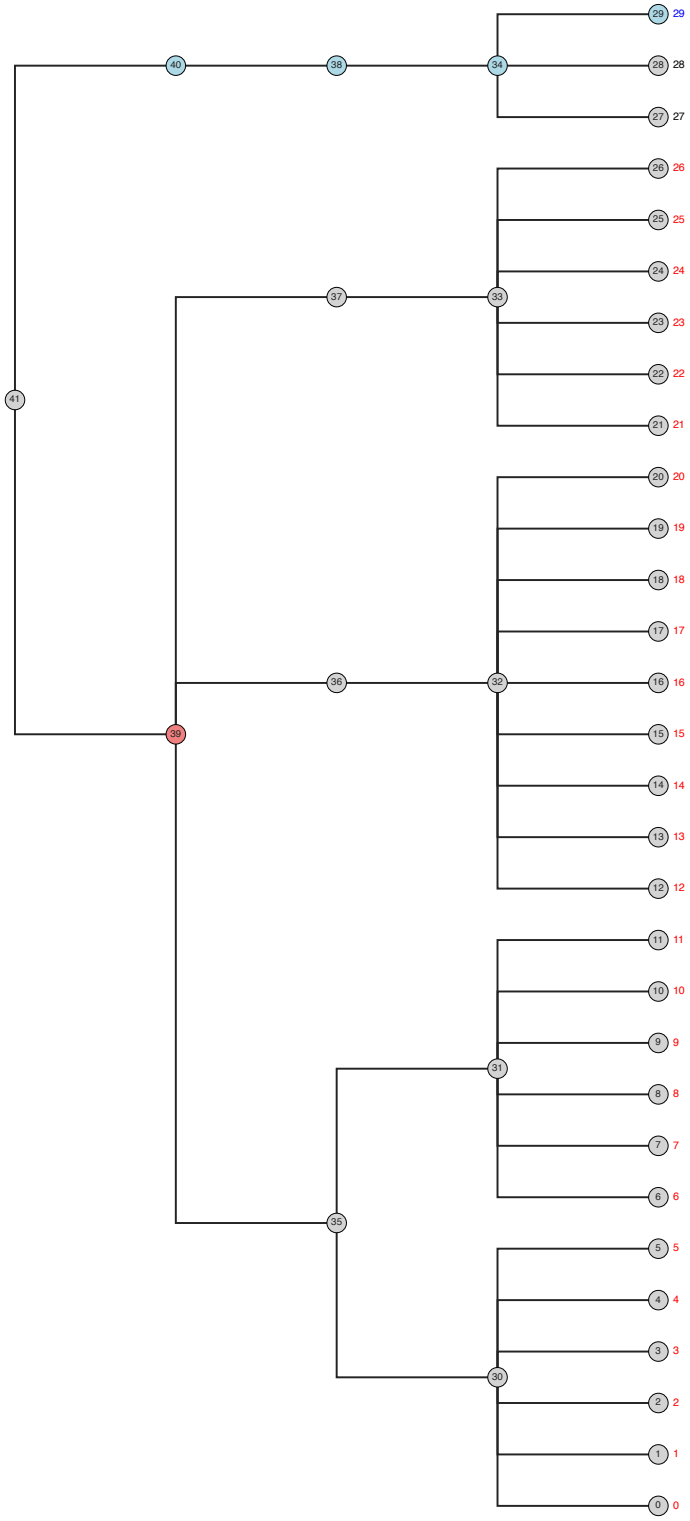


Figure S6. Randomly generated tree structure for synthetic data benchmark with one effect near the root of the tree, $p = 30$ tips. The red nodes were selected to be affected by the condition, causing the red tips to be differentially abundant. The blue tip is the reference feature, which forces the effects on all blue nodes to be 0.

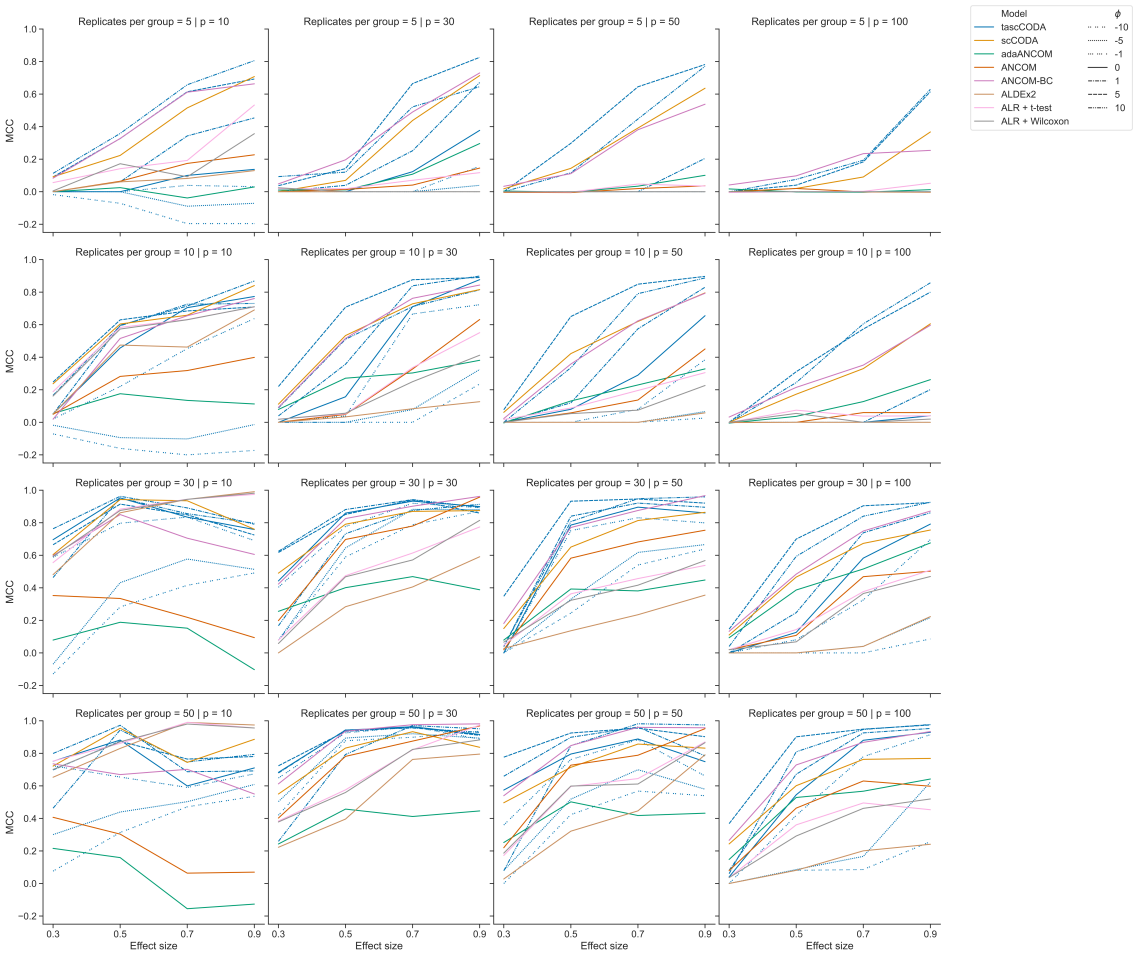


Figure S7. Matthews correlation coefficient (MCC) of tascCODA and other methods on simulated data with one binary covariate (differential abundance testing). Plots are grouped by the number of simulated components p, the number of samples per group and the effect size β. For tascCODA, different values of ϕ were tested (dashed blue lines).

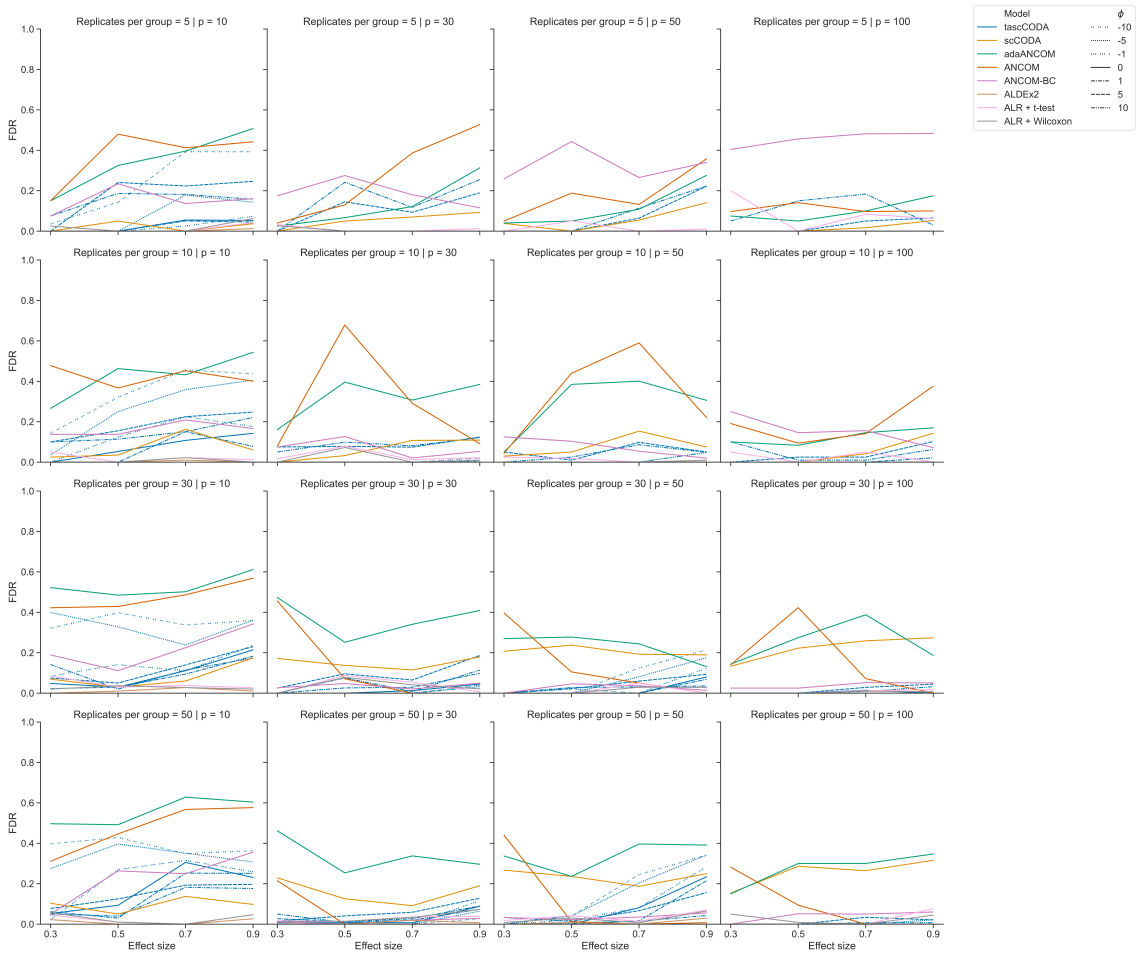


Figure S8. False discovery rate (FDR) of tascCODA and other methods on simulated data with one binary covariate (differential abundance testing). Plots are grouped by the number of simulated components p, the number of samples per group and the effect size β. For tascCODA, different values of ϕ were tested (dashed blue lines).

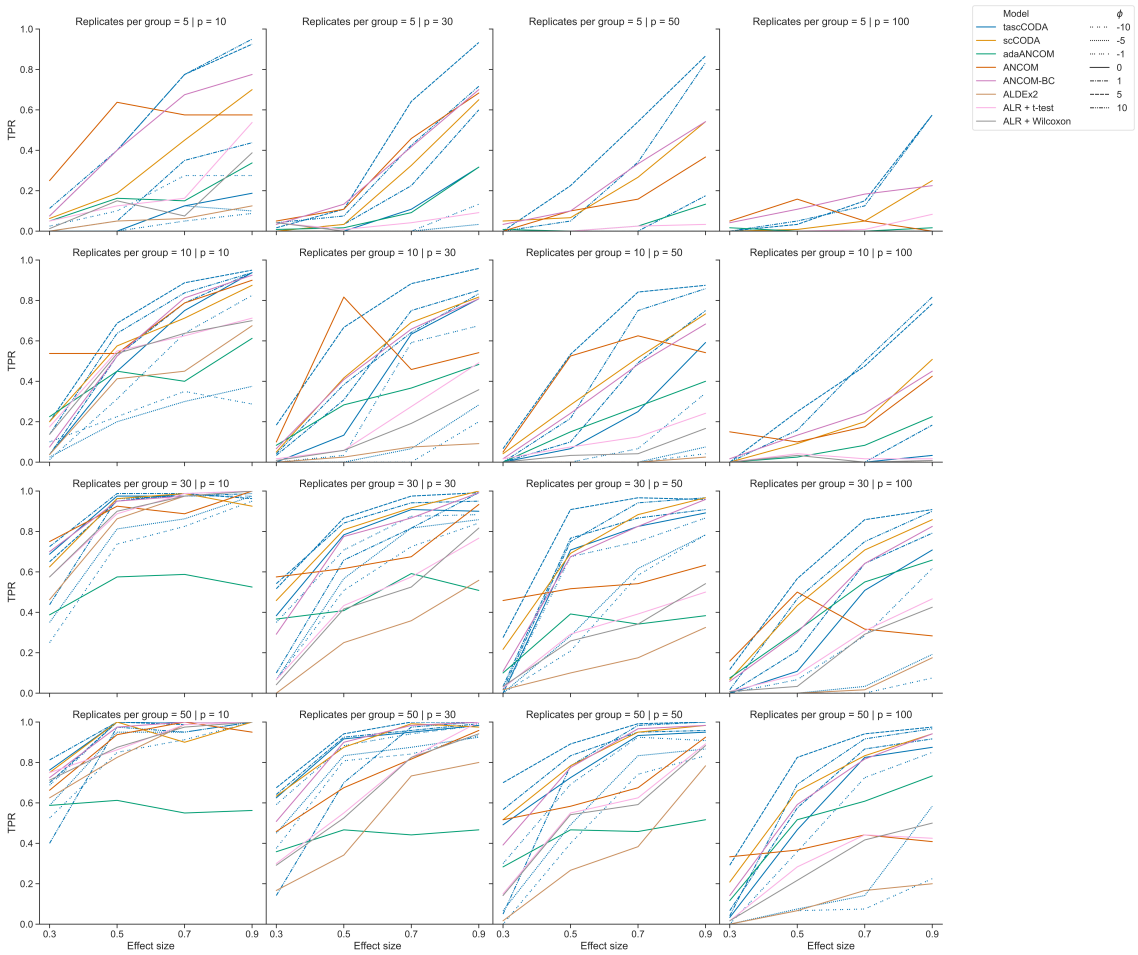


Figure S9. True positive rate (TPR) of tascCODA and other methods on simulated data with one binary covariate (differential abundance testing). Plots are grouped by the number of simulated components p, the number of samples per group and the effect size β. For tascCODA, different values of ϕ were tested (dashed blue lines).

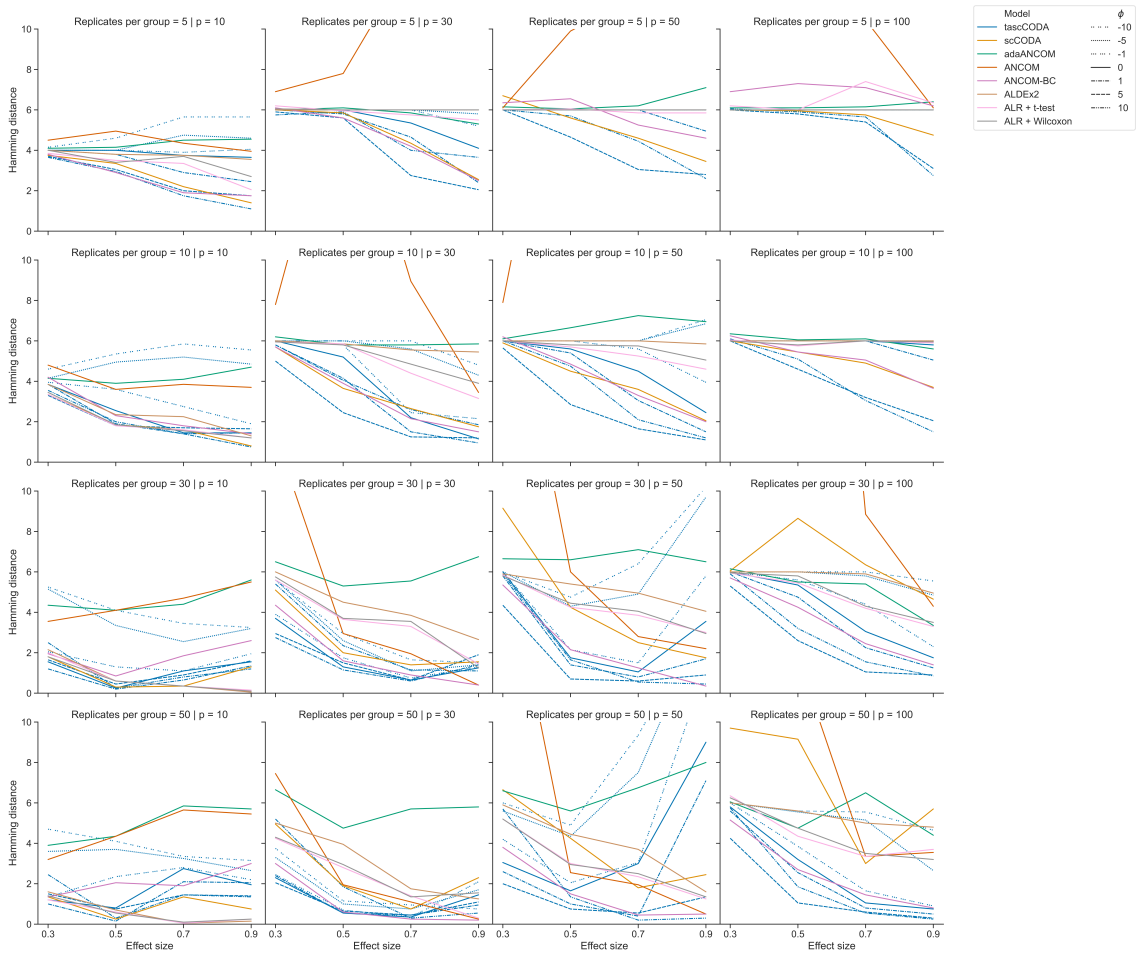


Figure S10. Hamming distance between ground truth and affected features determined by tascCODA and other methods on simulated data with one binary covariate (differential abundance testing). Plots are grouped by the number of simulated components p, the number of samples per group and the effect size β. For tascCODA, different values of ϕ were tested (dashed blue lines).

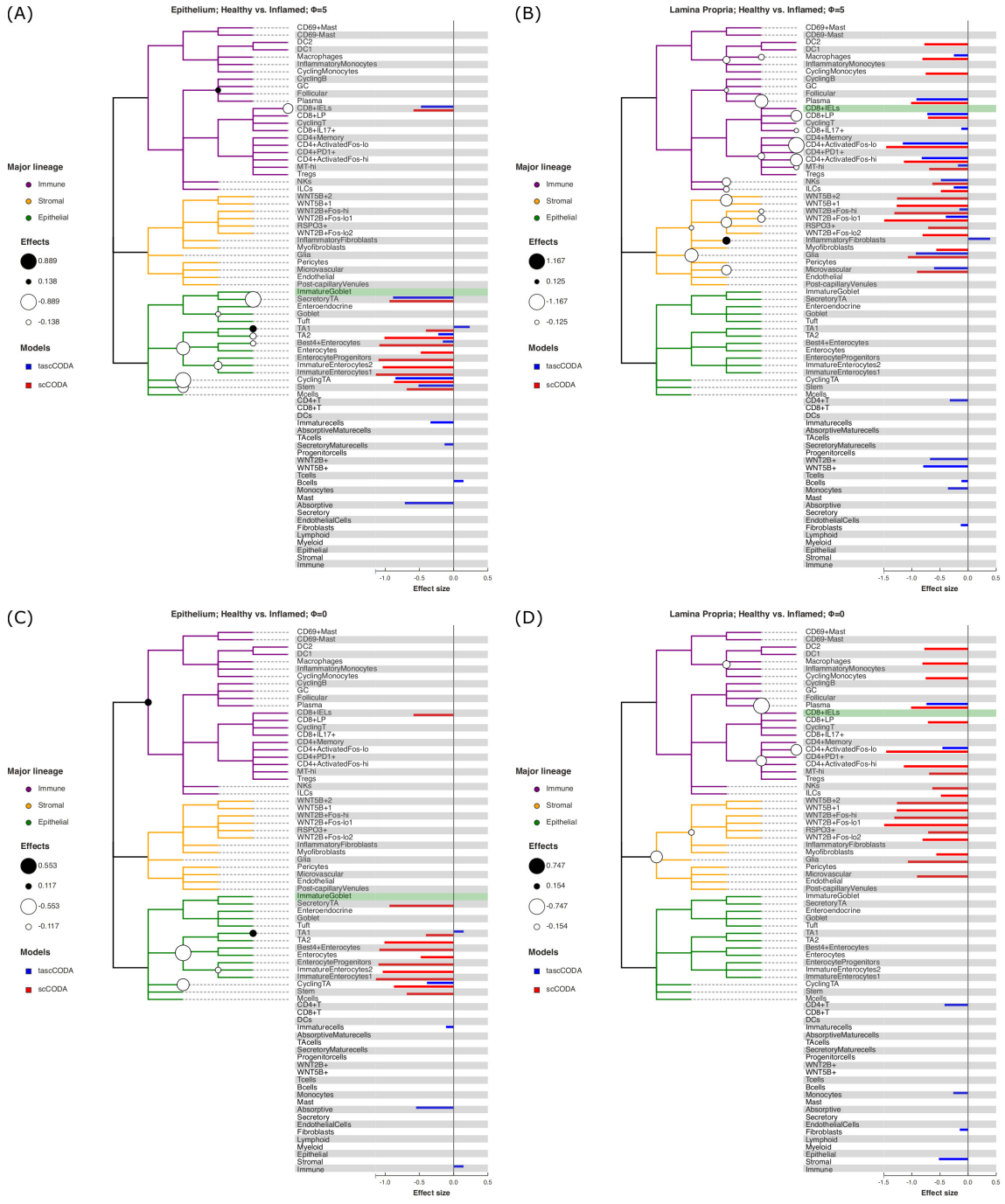


Figure S11. Behavior of tascCODA on scRNA-seq data. All plots show the case of comparing healthy control samples to inflamed tissue samples of UC patients in the data of Smillie et al. (2019). White and black circles on the cell lineage tree show the effects found by tascCODA, which are also shown as blue bars on the right side of each plot. The bars below the tree depict effects on internal nodes, with lower positions in the diagram corresponding to nodes closer to the root. For comparison, the red bars indicate effects found by scCODA, which only operates on the tips of the tree, on the same data. The green-shaded area shows the reference cell type that was used for both models. **(A)** $\phi = 5$, Epithelium. **(B)** $\phi = 5$, Lamina Propria. **(C)** $\phi = -3$, Epithelium. **(D)** $\phi = -3$, Lamina Propria.

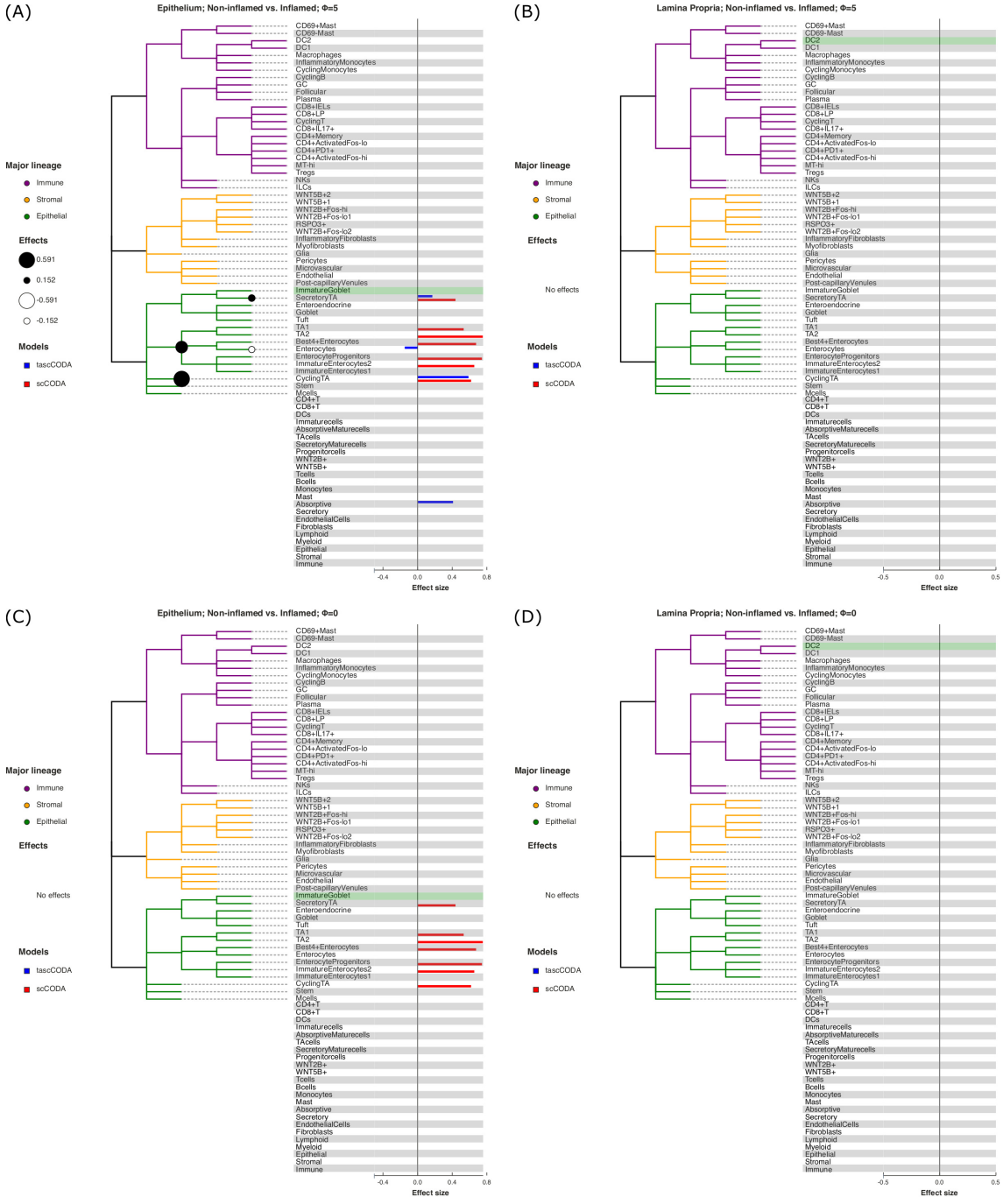


Figure S12. Behavior of tascCODA on scRNA-seq data. All plots show the case of non-inflamed to inflamed tissue samples of UC patients in the data of Smillie et al. (2019). White and black circles on the cell lineage tree show the effects found by tascCODA, which are also shown as blue bars on the right side of each plot. The bars below the tree depict effects on internal nodes, with lower positions in the diagram corresponding to nodes closer to the root. For comparison, the red bars indicate effects found by scCODA, which only operates on the tips of the tree, on the same data. The green-shaded area shows the reference cell type that was used for both models. **(A)** $\phi = 5$, Epithelium. **(B)** $\phi = 5$, Lamina Propria. **(C)** $\phi = -3$, Epithelium. **(D)** $\phi = -3$, Lamina Propria.

Table S1. Credible effects, highest density intervals, standard deviations and credibility threshold δ determined by tascCODA on epithelial biopsies from Smillie et al. (2019), $\phi = 5$. Abbreviations for scenarios: Healthy (H), Non-inflamed (N), and Inflamed (I).

Scenario	Node	Effect	HDI 3%	HDI 97%	SD	δ
Epithelium - H vs. N	ImmatureEnterocytes1	-0.647	-0.984	-0.309	0.181	0.148
	Enterocytes	-0.211	-0.637	0.054	0.211	0.148
	TA1	0.280	-0.022	0.475	0.134	0.148
Epithelium - H vs. I	Stem	-0.518	-1.000	0.004	0.286	0.135
	CyclingTA	-0.855	-1.144	-0.592	0.146	0.135
	Best4+Enterocytes	-0.163	-0.893	0.141	0.303	0.135
	TA2	-0.229	-0.802	0.129	0.282	0.135
	TA1	0.240	-0.159	0.730	0.265	0.135
	SecretoryTA	-0.889	-1.334	-0.477	0.228	0.135
	CD8+IELs	-0.481	-1.024	0.018	0.303	0.135
	Immaturecells	-0.343	-0.951	0.091	0.317	0.132
	SecretoryMaturecells	-0.138	-0.526	0.082	0.177	0.132
	Bcells	0.149	-0.068	0.525	0.178	0.130
	Absorptive	-0.717	-1.209	-0.110	0.299	0.126
Epithelium - N vs. I	CyclingTA	0.591	0.302	0.907	0.161	0.144
	Enterocytes	-0.152	-0.856	0.121	0.294	0.144
	SecretoryTA	0.174	-0.075	0.677	0.227	0.144
	Absorptive	0.413	-0.031	0.742	0.247	0.135

Table S2. Credible effects, highest density intervals, standard deviations and credibility threshold δ determined by tascCODA on Lamina Propria biopsies from Smillie et al. (2019), $\phi = 5$. Abbreviations for scenarios: Healthy (H), Non-inflamed (N), and Inflamed (I). For the N vs. I scenario, no credible effects were found.

Scenario	Node	Final Parameter	HDI 3%	HDI 97%	SD	Delta
LP - H vs. N	ImmatureGoblet	0.154	-0.161	0.787	0.270	0.131
	Microvascular	-0.354	-0.899	0.089	0.292	0.131
	Glia	-0.351	-0.867	0.083	0.278	0.131
	ILCs	-0.189	-0.710	0.125	0.242	0.131
	CD4+ActivatedFos-hi	-0.144	-0.544	0.100	0.184	0.131
	CD4+ActivatedFos-lo	-0.608	-1.048	-0.162	0.233	0.131
	CD8+LP	-0.169	-0.655	0.105	0.220	0.131
	Plasma	-0.472	-0.895	0.006	0.238	0.131
	TAcells	0.469	-0.038	0.952	0.281	0.129
	WNT2B+	-0.402	-0.772	0.043	0.245	0.126
	WNT5B+	-0.458	-0.935	0.067	0.296	0.129
	Tcells	-0.438	-0.778	0.043	0.263	0.117
	Bcells	-0.601	-1.047	-0.163	0.229	0.126
	Monocytes	-0.421	-0.817	0.044	0.258	0.124
LP - H vs. I	Microvascular	-0.612	-1.188	0.040	0.352	0.126
	Glia	-0.935	-1.558	-0.240	0.341	0.126
	InflammatoryFibroblasts	0.397	-0.155	1.425	0.481	0.126
	WNT2B+Fos-lo1	-0.403	-1.030	0.097	0.332	0.126
	WNT2B+Fos-hi	-0.160	-0.838	0.165	0.292	0.126
	ILCs	-0.261	-0.820	0.104	0.272	0.126
	NKs	-0.491	-0.964	0.025	0.277	0.126
	MT-hi	-0.186	-0.813	0.170	0.280	0.126
	CD4+ActivatedFos-hi	-0.830	-1.333	-0.310	0.272	0.126
	CD4+ActivatedFos-lo	-1.167	-1.686	-0.654	0.276	0.126
	CD8+IL17+	-0.127	-0.685	0.129	0.237	0.126
	CD8+LP	-0.732	-1.044	-0.409	0.168	0.126
	Plasma	-0.925	-1.217	-0.580	0.174	0.126
	Macrophages	-0.259	-0.811	0.097	0.266	0.126
	CD4+T	-0.331	-0.693	0.027	0.213	0.118
	WNT2B+	-0.683	-1.360	0.073	0.453	0.122
	WNT5B+	-0.803	-1.551	0.051	0.489	0.125
	Bcells	-0.125	-0.474	0.089	0.162	0.122
	Monocytes	-0.365	-0.848	0.068	0.283	0.120
	Fibroblasts	-0.136	-1.052	0.125	0.362	0.116

Table S3. Credible effects, highest density intervals, standard deviations and credibility threshold δ determined by tascCODA on biopsies from Smillie et al. (2019), $\phi = 0$. Abbreviations for scenarios: Healthy (H), Non-inflamed (N), and Inflamed (I). Credible effects were only found for one of six scenarios.

Scenario	Node	Effect	HDI 3%	HDI 97%	SD	δ
Epithelium - H vs. I	CyclingTA	-0.394	-0.669	0.010	0.193	0.074
	TA1	0.151	-0.023	0.496	0.176	0.074
	Immaturecells	-0.117	-0.500	0.026	0.177	0.074
	Absorptive	-0.553	-0.853	-0.205	0.179	0.074
	Immune	0.149	-0.015	0.324	0.108	0.074
LP - H vs. N	Plasma	-0.086	-0.524	0.037	0.185	0.066
	Tcells	-0.612	-0.796	-0.425	0.100	0.066
	Bcells	-0.761	-1.011	-0.380	0.173	0.066
	Monocytes	-0.315	-0.618	0.024	0.216	0.066
	Myeloid	-0.113	-0.511	0.035	0.184	0.066
	Epithelial	0.145	-0.013	0.322	0.106	0.066
	Stromal	-0.303	-0.483	0.007	0.143	0.066
LP - H vs. I	CD4+ActivatedFos-lo	-0.463	-0.967	0.034	0.316	0.063
	Plasma	-0.747	-0.963	-0.528	0.117	0.063
	CD4+T	-0.425	-0.708	-0.055	0.164	0.063
	Monocytes	-0.269	-0.568	0.019	0.197	0.063
	Fibroblasts	-0.154	-0.638	0.038	0.222	0.063
	Stromal	-0.525	-0.835	-0.148	0.184	0.063

Table S4. Credible effects found by tascCODA comparing the gut microbiome of healthy controls and IBS patients from Labus et al. (2017) for varying aggregation levels ϕ.

ϕ	Kingdom	Phylum	Class	Order	Family	Genus	Effect
-5	Bacteria	Firmicutes					-0.313
0	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Tannerellaceae	Parabacteroides	-0.156
0	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	-0.662
0	Bacteria	Firmicutes	Clostridia	Oscillospirales			-0.232
5	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Tannerellaceae	Parabacteroides	-0.845
5	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	-1.001
5	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella	-0.413
5	Bacteria	Firmicutes	Clostridia	Lachnospirales	Lachnospiraceae	Agathobacter	-0.610
5	Bacteria	Firmicutes	Clostridia	Oscillospirales	Ruminococcaceae	Subdoligranulum	-0.224
5	Bacteria	Firmicutes	Clostridia	Oscillospirales	Ruminococcaceae	Faecalibacterium	-0.252
5	Bacteria	Firmicutes	Negativicutes	Acidaminococcales	Acidaminococcaceae	Phascolarctobacterium	-0.250
5	Bacteria	Firmicutes	Clostridia	Oscillospirales	Ruminococcaceae		-0.340

Table S5. Credible effects found by tascCODA ($\phi = 5$) comparing the gut microbiome of four different subtypes of IBS to all other samples. Original data by Labus et al. (2017).

Subtype	Kingdom	Phylum	Class	Order	Family	Genus	Effect
IBS-C	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	-0.426
IBS-C	Bacteria	Firmicutes	Clostridia	Lachnospirales	Lachnospiraceae	Anaerostipes	0.438
IBS-C	Bacteria	Firmicutes	Clostridia	Lachnospirales	Lachnospiraceae	Agathobacter	-0.819
IBS-C	Bacteria	Firmicutes	Clostridia	Oscillospirales	Ruminococcaceae	Ruminococcus	-0.262
IBS-C	Bacteria	Firmicutes	Clostridia	Oscillospirales	Ruminococcaceae	Faecalibacterium	-0.320
IBS-D	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Tannerellaceae	Parabacteroides	-0.392
IBS-D	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	-1.405
IBS-M	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Tannerellaceae	Parabacteroides	-0.424
IBS-M	Bacteria	Firmicutes	Clostridia	Lachnospirales	Lachnospiraceae	Blautia	0.799
IBS-M	Bacteria	Firmicutes	Clostridia	Oscillospirales	Ruminococcaceae	Faecalibacterium	-0.285
IBS-unspecified	Bacteria	Firmicutes	Clostridia	Peptostreptococcales-Tissierellales	Peptostreptococcaceae	Romboutsia	0.259

REFERENCES

Bai, R., Rockova, V., and George, E. I. (2020). Spike-and-Slab meets LASSO: A review of the Spike-and-Slab LASSO

Büttner, M., Ostner, J., Müller, C. L., Theis, F. J., and Schubert, B. (2020). scCODA: A bayesian model for compositional single-cell data analysis. doi:10.1101/2020.12.14.422688

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: High-resolution sample inference from illumina amplicon data. *Nat. Methods* 13, 581–583. doi:10.1038/nmeth.3869

Labus, J. S., Hollister, E. B., Jacobs, J., Kirbach, K., Oezguen, N., Gupta, A., et al. (2017). Differences in gut microbial composition correlate with regional brain volumes in irritable bowel syndrome. *Microbiome* 5, 49. doi:10.1186/s40168-017-0260-z

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–6. doi:10.1093/nar/gks1219

Ročková, V. and George, E. I. (2018). The Spike-and-Slab LASSO. *J. Am. Stat. Assoc.* 113, 431–444. doi:10.1080/01621459.2016.1260469

Smillie, C. S., Biton, M., Ordovas-Montanes, J., Sullivan, K. M., Burgin, G., Graham, D. B., et al. (2019). Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* 178, 714–730.e22. doi:10.1016/j.cell.2019.06.029

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* 58, 267–288

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., et al. (2014). The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* 42, D643–8. doi:10.1093/nar/gkt1209

B. Score matching for differential abundance testing of compositional high-throughput sequencing data

Contributing article

Ostner, J., Li, H., and Müller, C.L. (2024). Score matching for differential abundance testing of compositional high-throughput sequencing data. *bioRxiv, 2024-12.* doi: `https://doi.org/10.1101/2024.12.05.627006`

Replication code

Source code for this contribution has been deposited on Github at `https://github.com/bio-datascience/cosmoDA`. Supplemental data can be downloaded from zenodo (`https://zenodo.org/records/13911623`).

Copyright information

Author contributions

J.O. developed the idea of cosmoDA with help from H.L. and C.L.M. J.O. implemented the method, performed and evaluated all simulations and conducted the data applications. J.O. wrote the manuscript with assistance from C.L.M. All authors read and approved the final manuscript.

Score matching for differential abundance testing of compositional high-throughput sequencing data

Johannes Ostner[1,2*], Hongzhe Li[3] and Christian L. Müller[1,2,4]

[1]Computational Health Center, Helmholtz Munich, Neuherberg, Germany.
[2]Institut für Statistik, Ludwig-Maximilians-Universität München, Munich, Germany.
[3]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.
[4]Center for Computational Mathematics, Flatiron Institute, New York, NY, USA.

Abstract

The class of a-b power interaction models, proposed by Yu et al. (2024), provides a general framework for modeling sparse compositional count data with pairwise feature interactions. This class includes many distributions as special cases and enables zero count handling through power transformations, making it especially suitable for modern high- throughput sequencing data with excess zeros, including single-cell RNA-Seq and amplicon sequencing data. Here, we present an extension of this class of models that can include covariate information, allowing for accurate characterization of covariate dependencies in heterogeneous populations. Combining this model with a tailored differential abundance (DA) test leads to a novel DA testing scheme, `cosmoDA`, that can reduce false positive detection caused by correlated features. `cosmoDA` uses the generalized score matching estimation framework for power interaction models Our benchmarks on simulated and real data show that `cosmoDA` can accurately estimate feature interactions in the presence of population heterogeneity and significantly reduces the false discovery rate when testing for differential abundance of correlated features. Finally, `cosmoDA` provides an explicit link to popular Box-Cox-type data transformations and allows to assess the impact of zero replacement and power transformations on downstream differential abundance results. `cosmoDA` is available at https://github.com/bio-datascience/cosmoDA.

Keywords: Compositional data, Score matching, Differential abundance, Generative model, Single-cell RNA sequencing, Microbiome

1 Introduction

Count matrices, detailing the compositional makeup of cellular constituents in a sample, are an important data modality derived from modern high-throughput sequencing (HTS) experiments, including amplicon sequencing (Quinn et al., 2018; Tsilimigras and Fodor, 2016) and single-cell RNA-Sequencing (scRNA-Seq) (Büttner et al., 2021; Heumos et al., 2023). These matrices commonly have the form $\tilde{\mathbf{X}} \in \mathbb{N}_0^{n \times p}$ and show the abundance of p features (cell types or microbial taxa) in n tissues (Regev et al., 2017), bacterial communities (McNulty et al., 2023), or microbial habitats (Turnbaugh et al., 2007). Because sequencing capacity of HTS experiments is technically limited, each sample only represents a small part of a larger population, rendering the sum of counts in a row non-quantitative and making the data compositional (Gloor et al., 2017). Dividing each sample by its total sum yields relative abundance data, which is proportionally equivalent to the original data and constrained to the $(p-1)$-dimensional probability simplex (Aitchison, 1982):

$$\Delta \equiv \Delta_{p-1} = \left\{ \boldsymbol{x} \in \mathbb{R}^p : \boldsymbol{x} \succeq \mathbf{0},\, \mathbf{1}_p^\top \boldsymbol{x} = 1 \right\}. \tag{1}$$

Generative models for HTS-derived compositional data commonly respect compositionality either by transforming the data into Euclidean space through log-ratio or similar transformations (Love et al., 2014; Mishra and Müller, 2022), or by using distributions directly defined on the probability simplex. The Dirichlet distribution is a popular choice due to its relatively simple structure and interpretability (Hijazi and Jernigan, 2009; Wadsworth et al., 2017; Büttner et al., 2021; Ostner et al., 2021). The assumption of independent features (apart from the compositional effect) is, however, a major limitation of the Dirichlet distribution. To allow for more complex dependency structures, Aitchison and Shen (1980) proposed the class of logistic normal distributions, which include the estimation of feature-feature interactions. Several lines of research make successful use of logistic normal models (and extensions thereof) for HTS data (see, e.g., Xia et al. (2013); Zeng et al. (2022)) and address the computational challenges in scaling parameter inference to large-scale datasets (Silverman et al., 2022).

Another challenge in generative HTS data modeling is the presence of zeroes. Since the logistic normal distribution requires the underlying data to be positive due to logarithmic transformations, zero entries in the primary data need to be replaced by positive values (Lubbe et al., 2021; Greenacre et al., 2023). Any such procedure inevitably distorts the measured data compositions, especially for rare features with many zero entries (Lubbe et al., 2021), resulting in another source of modeling inaccuracy.

In his seminal work, Aitchison (1985) provided a general class of distribution, the A^{p-1} class, that includes the logistic normal and the Dirichlet distribution as special case. This class forms the basis for more recent models that extend the A^{p-1} class and do not require zero imputation. For low-dimensional data, Scealy and Wood (2022) and Scealy et al. (2024) introduced the polynomially tilted pairwise interaction (PPI) model, which has properties similar to the Dirichlet distribution at the boundaries of the simplex. The class of a-b power interaction models, introduced by Yu et al. (2024), achieves validity on the simplex boundaries by replacing the logarithm with power

transformations. These works further use score matching estimation (Hyvärinen, 2005) for computationally efficient parameter inference, reducing the estimation problem to solving a (regularized) quadratic optimization problem. However, both the PPI and the a-b power interaction model currently only allow to model a homogeneous sample population and cannot describe differences between groups of samples.

A central task in HTS data analysis is the detection of significant differences in the feature composition, given environmental, clinical, or host-specific perturbations or variations. This problem, also known as differential abundance (DA) testing, faces the same challenges as generative modeling (Gloor et al., 2017; Tsilimigras and Fodor, 2016). While compositionality and zero handling are discussed in most state-of-the-art DA testing methods (Lin and Peddada, 2020; Zhou et al., 2022; Nearing et al., 2022), only few methods explicitly include interactions between compositional features in their testing procedure (Ma et al., 2024). Such interactions, however, may contribute to the false discovery of certain features that are not directly impacted by the perturbations or covariate changes, but simply strongly correlate with the differentially abundant feature. Consider a composition of five microbial taxa a, b, c, d, e, where a and b have a symbiotic relationship and their abundances are highly correlated (Figure 1a). A treatment now targets taxon a and causes a decline in its population. This will in turn cause the abundance of taxon b to also decrease, although it was not directly influenced by the treatment. Classical DA testing methods will not be able to discern between these primary and secondary effects caused by the treatment, detecting both a and b as differentially abundant.

In this work, we present a new DA framework, termed `cosmoDA` (**co**mpositional **s**core **m**atching **o**ptimization for **D**ifferential **A**bundance analysis), that addresses the challenge of feature interactions in DA testing. `cosmoDA` is based on the a-b power interaction models from Yu et al. (2024) and introduces a linear covariate effect on the location vector, thus enabling the inclusion of sample group indicators or continuous covariates of interest. We provide a framework for assessing the significance of the estimated covariate effects, which, in the case of group indicator variables, allows principled compositional differential abundance testing. A similar covariate-extended model was introduced for low-dimensional compositional data by Billheimer et al. (2001), albeit only for the special case of the logistic normal model. In the a-b power interaction models, maximum likelihood estimation is not possible due to the intractability of computing the normalzing constant. We thus resort to the score matching framework (Hyvärinen, 2005). By carefully studying the structural properties of the underlying score matching objective, our extended estimation framework retains the favorable quadratic nature of the underlying optimization problem with negligible computational overhead. Regularization on the interaction effects further ensures model identifiability and selection of the most important correlation patterns. The characteristics of the a-b power interaction model thus ensure that feature interactions are adequately considered and zero entries in the underlying data do not need to be replaced or imputed.

The remainder of the paper is structured as follows. In the next section, we introduce `cosmoDA` as an extension of the a-b power interaction model, describe the score matching estimation framework, and explain how the power transformation makes zero

replacement obsolete. We then describe the model regularization framework, sketch the computational implementation, and introduce the differential abundance testing framework `cosmoDA`. Section 3 provides several simulated data benchmarks that showcase the ability of `cosmoDA` to (i) correctly estimate sparse interaction matrices in the presence of covariates and (ii) reduce the false discovery rate in differential abundance testing compared to other state-of-the-art methods. We investigate the impact of different power transformations on differential abundance in real scRNA-seq and 16S rRNA sequencing data in section 4 and provide a data-driven method to select the power exponents in practice. Section 5 discusses the results, highlights strengths and limitations of the work, and provides guidelines for future research. `cosmoDA` is available as a Python package at https://github.com/bio-datascience/cosmoDA.

2 Methods

We consider to model compositional matrices $\boldsymbol{X}$ where each row $\boldsymbol{x}^{(i)} \in \Delta$, $i = 1, \ldots, n$, represents a sample, and each column $\boldsymbol{x}_j$, $j = 1, \ldots, p$, represents the jth compositional feature. We are motivated by the large number of available biological "compositional" count matrices $\tilde{\boldsymbol{X}} \in \mathbb{N}_0^{n \times p}$ derived from HTS experiments. Important instances include 16S rRNA amplicon sequencing data, where each feature represents read counts associated with a microbial taxon Gloor et al. (2017) and scRNA-Seq experiments, where each feature represents a certain cell-type proportion, as derived from clustered transcriptional profiles (Büttner et al., 2021; Heumos et al., 2023). Due to the compositional nature of the derived count data, a common approach is to scale each observation $\tilde{\boldsymbol{x}}^{(i)}$ by its library size $S^{(i)} = \sum_{j=1}^{p} x_j^{(i)}$ to obtain relative abundance samples $\boldsymbol{x}^{(i)} = \tilde{\boldsymbol{x}}^{(i)} / S^{(i)} \in \Delta$ (Gloor et al., 2017).

2.1 The covariate-extended a-b power interaction model

Following the proposal in Yu et al. (2024), we model samples in $\boldsymbol{X}$ through the *a-b power interaction model* on the $(p-1)$-dimensional simplex Δ. The unnormalized probability density for one sample $\boldsymbol{x} \equiv \boldsymbol{x}^{(i)}$ reads:

$$p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^{a} + \frac{1}{b}\boldsymbol{\eta}^{\top}\boldsymbol{x}^{b}\right), \quad \boldsymbol{x} \in \Delta;\ \boldsymbol{K} = \boldsymbol{K}^{T};\ \boldsymbol{K}\mathbf{1}_p = \mathbf{0}_p\,. \tag{2}$$

Here, interactions between features are modeled through the *interaction matrix* $\boldsymbol{K} \in \mathbb{R}^{p \times p}$, and the *location vector* $\boldsymbol{\eta} \in \mathbb{R}^p$ describes the base composition of the individual features. This model belongs to the class of exponential family models. Using the conventions in Yu et al. (2024), $\boldsymbol{x}^a \equiv \log(\boldsymbol{x}); 1/a \equiv 1$ if $a = 0$, and $\boldsymbol{x}^b \equiv \log(\boldsymbol{x}); 1/b \equiv 1$ if $b = 0$, power interaction models encapsulate several probability distributions as special cases.

With parameter settings $a = b = 0$, the model includes the Dirichlet distribution with the additional constraints $\boldsymbol{K} = 0, \eta \succ -1$, the logistic normal distribution (Aitchison and Shen, 1980) with the constraints $\boldsymbol{K}\mathbf{1}_p = \mathbf{0}_p, \boldsymbol{x}^T\boldsymbol{K}\boldsymbol{x} > 0\ \forall \boldsymbol{x}, \mathbf{1}_p^T\boldsymbol{\eta} = -p$, and Aitchison's A^{p-1} family of distributions Aitchison (1985) with $\boldsymbol{x}^T\boldsymbol{K}\boldsymbol{x} > 0\ \forall \boldsymbol{x}, \boldsymbol{\eta} \succeq -1$. For the logistic normal case, the interaction matrix $\boldsymbol{K}$ is equivalent to the inverse covariance matrix of logratio-transformed data, given specific linear transformations (Erb, 2020). With parameter settings $a = 1$ and $b = 0$, the model is equivalent to the PPI distribution (Scealy and Wood, 2022; Scealy et al., 2024) (see Appendix A), and with parameter settings $a = b = 1$, the power interaction model is equivalent to the maximum entropy distribution on the simplex with the constraints $\boldsymbol{K}\mathbf{1}_p = \mathbf{0}_p, \boldsymbol{x}^T\boldsymbol{K}\boldsymbol{x} > 0\ \forall \boldsymbol{x}$, as derived in Weistuch et al. (2022).

As stated in Theorem 1 from Yu et al. (2024), the probability density in Eq. (2) is proper if either

- $a > 0,\ b > 0$;
- $a > 0,\ b = 0,\ \eta_j > -1\ \forall j$;
- $a = 0,\ b = 0,\ \log(\boldsymbol{x})^T\boldsymbol{K}\log(\boldsymbol{x}) > 0\ \forall \boldsymbol{x} \in \Delta$;

- $a = 0,\ b > 0,\ \log(\boldsymbol{x})^T \boldsymbol{K} \log(\boldsymbol{x}) \geq 0\ \forall \boldsymbol{x} \in \Delta$.

We next extend the original proposal of the a-b power interaction model by including a (continuous or binary) covariate vector $\boldsymbol{y} \in \mathbb{R}^n$ (or $\boldsymbol{y} \in \{0,1\}^n$, respectively) in the model. The covariate describes, e.g., a concurrently measured quantity of interest for each sample, or, more relevant in our context, a condition-specific indicator vector. We model the influence of $y \equiv y^{(i)}$ on $\mathbf{x}^{(i)}$ by introducing a linear model on the location vector $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} = \boldsymbol{\eta}_0 + y\boldsymbol{\eta}_1\,. \tag{3}$$

Plugging this model into Eq. (2) results in the covariate-extended a-b power interaction model:

$$p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^a + \frac{1}{b}(\boldsymbol{\eta}_0 + y\boldsymbol{\eta}_1)^\top \boldsymbol{x}^b\right), \quad \boldsymbol{x} \in \Delta;\ \boldsymbol{K}\mathbf{1}_p = \boldsymbol{K}^T\mathbf{1}_p = \mathbf{0}_p\,. \tag{4}$$

This formulation of the model assumes that all samples stem from an overall population with fixed interaction matrix $\boldsymbol{K}$, but allows the proportions of features, described by $\boldsymbol{\eta}$, to be dependent on the measured covariate y. For the probability density of the covariate-extended a-b power interaction model to be proper, the same conditions hold as for the model in Eq. (2), replacing $\boldsymbol{\eta}$ with $\boldsymbol{\eta}_0 + y\boldsymbol{\eta}_1$.

The model in Eq. (4) forms the basis for our differential abundance testing framework `cosmoDA` (**co**mpositional **s**core **m**atching **o**ptimization for **D**ifferential **A**bundance analysis). In case where $\boldsymbol{y}$ represents a binary group indicator, e.g., case vs. control samples, `cosmoDA` fits the data to the model and tests for significant changes of the individual components of $\boldsymbol{\eta}_1$. Figure 1 provides a conceptual overview of `cosmoDA`. Before detailing the specific test statistics, we describe the underlying parameter estimation framework.

2.2 Model estimation

2.2.1 Score matching for power interaction models

Efficient parameter estimation for the a-b power interaction models (Eq. 2) through generalized score matching (Hyvärinen, 2005, 2007; Yu et al., 2019) was proposed by Yu et al. (2024). Given an (unknown) true data distribution P_0 with density p_0 and a family of distributions of interest $\mathcal{P}(\mathcal{D})$, score matching tries to find $P \in \mathcal{P}(\mathcal{D})$ with density p such that the Hyvärinen divergence between the gradients of the logarithm of the densities of P_0 and P is minimized:

$$\frac{1}{2}\int_{\mathcal{D}} p_0(\boldsymbol{x}) \left\| \nabla \log p(\boldsymbol{x}) \odot \tilde{\boldsymbol{h}}^{1/2}(\boldsymbol{x}) - \nabla \log p_0(\boldsymbol{x}) \odot \tilde{\boldsymbol{h}}^{1/2}(\boldsymbol{x}) \right\|_2^2 \mathrm{d}\boldsymbol{x}\,, \tag{5}$$

where $\tilde{\boldsymbol{h}}(\boldsymbol{x}) = (\tilde{h}_1(x_1), \ldots, \tilde{h}_p(x_p))$ is a weight function. Yu et al. (2022) show that score matching can be performed on domains with positive Lebesgue measure in $\mathbb{R}^p$ by setting $\tilde{h}$ such that $\nabla \log p(\boldsymbol{x}) \odot \tilde{\boldsymbol{h}}^{1/2}(\boldsymbol{x})$ does not vanish at the boundaries of the domain.

Yu et al. (2024) adapted the generalized score matching framework for the a-b power interaction models on the $(p-1)$-dimensional probability simplex in $\mathbb{R}^p$ by

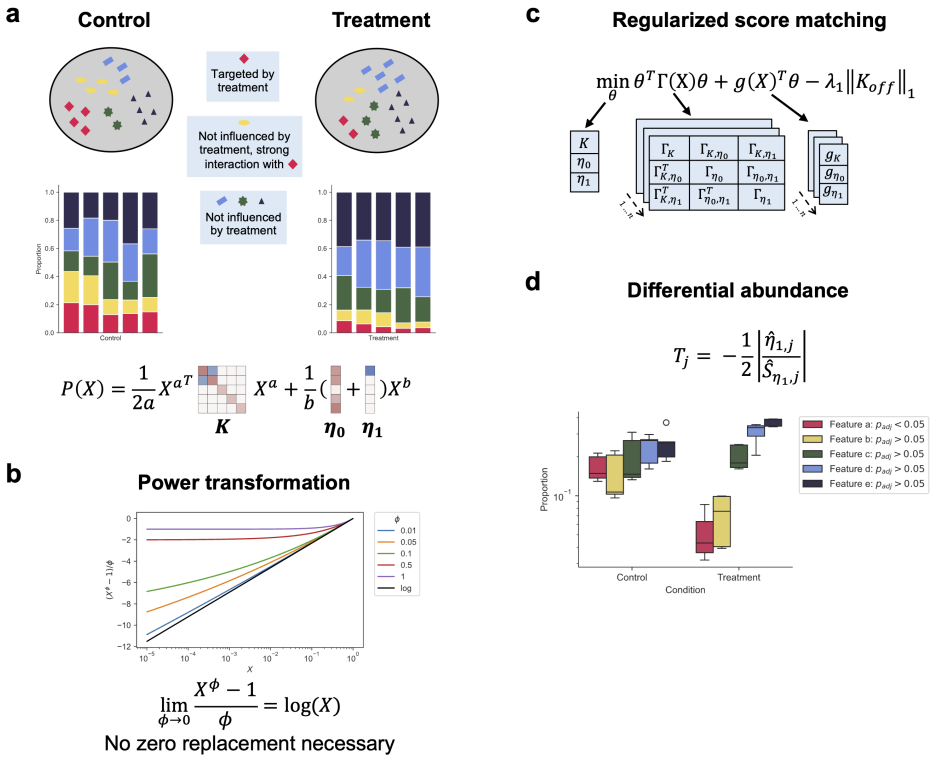


Fig. 1 `cosmoDA` allows to perform generative modeling and differential abundance testing on compositional data with feature interactions. (a) Interactions between features can alter the abundance of features although they are not directly associated with the condition. `cosmoDA` is able to accurately distinguish primary from secondary effects by inferring pairwise feature interactions in addition to the effects associated with the condition. (b) Power transformations allow to analyze compositional data without imputation of zero values. For decreasing exponents, the Box-Cox transformation converges to the logarithm. (c) `cosmoDA` uses regularized score matching for parameter inference. The optimization problem therefore reduces to a quadratic function with parameters Γ and g defined by averaging over all samples. (d) Differential abundance testing in `cosmoDA` uses a studentized test statistic. Only the feature primarily associated with the condition (Feature a) has a small adjusted p-value.

profiling out the last coordinate $x_p \equiv 1 - \sum_{j=1}^{m-1} x_j$, similar to the additive log-ratio transformation. We follow this approach, setting $\tilde{h}_j(\boldsymbol{x}) = (h_j \circ \varphi_j)(\boldsymbol{x})$ with $h_j(\boldsymbol{x}) = x_j^c$ and $\varphi_j(\boldsymbol{x}) = \min\{x_j, x_p, C_j\}$ and fixing $C_j = 1$ and $c = 2$, as recommended by Yu et al. (2024). This results in a weight function $\tilde{h}_j(\boldsymbol{x}) = \min\{x_j, x_p\}^2$. With $p(\boldsymbol{x})$ from the family of a-b power interaction models, the following mild assumptions hold (Yu et al., 2024):

1. $p_0(x_j; \boldsymbol{x}_{-j}) h_j(\varphi(\boldsymbol{x})_j) \partial_j \log p(x_j; \boldsymbol{x}_{-j}) \Big|_{x_j \nearrow b_k(\boldsymbol{x}_{-j})^-}^{x_j \searrow a_k(\boldsymbol{x}_{-j})^+} = 0$
   for all $k = 1, \ldots, K_j(\boldsymbol{x}_{-j})$ and $\boldsymbol{x}_{-j} \in \mathcal{S}_{-j,\mathcal{D}}$ for all $j = 1, \ldots, p$;

2. $\int_{\mathcal{D}} p_0(\boldsymbol{x}) \left\| \nabla \log p(\boldsymbol{x}) \odot (\boldsymbol{h} \circ \boldsymbol{\varphi})^{1/2}(\boldsymbol{x}) \right\|_2^2 \mathrm{d}\boldsymbol{x} < +\infty$,
   $\int_{\mathcal{D}} p_0(\boldsymbol{x}) \left\| [\nabla \log p(\boldsymbol{x}) \odot (\boldsymbol{h} \circ \boldsymbol{\varphi})(\boldsymbol{x})]' \right\|_1 \mathrm{d}\boldsymbol{x} < +\infty$.
3. $\forall j = 1, \ldots, p$ and almost everywhere $\boldsymbol{x}_{-j} \in \mathcal{S}_{-j,\mathcal{D}},$, the component function h_j of $\boldsymbol{h}$ is absolutely continuous in any bounded sub-interval of the section $\mathcal{C}_{j,\mathcal{D}}(\boldsymbol{x}_{-j})$.

Therefore, a consistent estimator of the loss function (Eq. 5) follows as a sample- and feature-wise sum over the entire dataset:

$$\hat{L}_{\boldsymbol{h}}(P) = \frac{1}{2} \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{1}{2} (h_j \circ \varphi_j) \left(\boldsymbol{X}^{(i)} \right) \cdot \left[\partial_j \log p \left(\boldsymbol{X}^{(i)} \right) \right]^2 + \partial_j \left[(h_j \circ \varphi_j) \left(\boldsymbol{X}^{(i)} \right) \cdot \partial_j \log p \left(\boldsymbol{X}^{(i)} \right) \right], \quad (6)$$

where $\boldsymbol{X}^{(i)}$, $1 \leq i \leq n$, form an i.i.d. sample from the unknown data distribution P_0 and P is an a-b power interaction model with unnormalized density as described in Eq. (2). Aggregating $\boldsymbol{K}$ and $\boldsymbol{\eta}$ to $\boldsymbol{\theta} = (\mathrm{vec}(\boldsymbol{K}), \boldsymbol{\eta})$ and defining $P_{\boldsymbol{\theta}}$ and its density $p_{\boldsymbol{\theta}}$ accordingly shows that the power interaction model without covariate (Eq. 2) follows an exponential-family-type model

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{t}(\boldsymbol{x}) - \psi(\boldsymbol{\theta}) + b(\boldsymbol{x}) \,, \boldsymbol{x} \in \Delta \,, \quad (7)$$

where the function $\boldsymbol{t}(\cdot)$ denotes the function for the sufficient statistics, $\psi(\cdot)$ the cumulant function, and $b(\cdot)$ the logarithm of the base measure, respectively.

Then, $\hat{L}_{\boldsymbol{h}}$ can be reformulated as a quadratic optimization problem:

$$\hat{L}_{\boldsymbol{h}}(P_{\boldsymbol{\theta}}) = \frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\Gamma}(\boldsymbol{x}) \boldsymbol{\theta} - \boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{\theta} + \mathrm{const.} \quad (8)$$

with $\boldsymbol{\Gamma}(\mathbf{x}) \in \mathbb{R}^{r \times r}$ and $\boldsymbol{g}(\boldsymbol{x}) \in \mathbb{R}^r$ are sample averages of known functions in $\boldsymbol{x}$ only. Analogously, the same considerations hold for the covariate-extended a-b power interaction model (Eq. 4), substituting $\boldsymbol{\eta}$ with $\boldsymbol{\eta}_0 + y\boldsymbol{\eta}_1$. This substitution does not yet provide individual estimates of $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_0$ though, which are required for differential abundance testing. To obtain these individual estimates, a look at the exact derivation of $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$, as described by Yu et al. (2024), is necessary. We first split the location vector into its two parts $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$, and set $\boldsymbol{\theta} = (\mathrm{vec}(\boldsymbol{K}), \boldsymbol{\eta_0}, \boldsymbol{\eta_1})$. After dropping the last coordinate by assuming $x_p \equiv 1 - \mathbf{1}_{p-1}^\top \boldsymbol{x}_{-p}$ as above, the first and second partial derivatives for the covariate-less model (Eq. 4.1 and 4.2 in Yu et al. (2024)) can easily be adapted to the covariate-extended model:

$$\begin{aligned}
\partial_j \log p(\boldsymbol{x}_{-p}) &= -\left(\boldsymbol{\kappa}_{,j}^\top \boldsymbol{x}^a\right) x_j^{a-1} + \left(\boldsymbol{\kappa}_{,p}^\top \boldsymbol{x}^a\right) x_p^{a-1} + \eta_j x_j^{b-1} - \eta_p x_p^{b-1}, && (9)\\
&= -\left(\boldsymbol{\kappa}_{,j}^\top \boldsymbol{x}^a\right) x_j^{a-1} + \left(\boldsymbol{\kappa}_{,p}^\top \boldsymbol{x}^a\right) x_p^{a-1} \\
&\quad + \eta_{0,j} x_j^{b-1} - \eta_{0,p} x_p^{b-1} + y_j \eta_{1,j} x_j^{b-1} - y_p \eta_{1,p} x_p^{b-1} \\
\partial_{jj} \log p(\boldsymbol{x}_{-p}) &= -(a-1)\left[\left(\boldsymbol{\kappa}_{,j}^\top \boldsymbol{x}^a\right) x_j^{a-2} + \left(\boldsymbol{\kappa}_{,p}^\top \boldsymbol{x}^a\right) x_p^{a-2}\right] && (10)\\
&\quad - a\left[\kappa_{jj} x_j^{2a-2} + \kappa_{pp} x_p^{2a-2} + 2\kappa_{jp} x_j^{a-1} x_p^{a-1}\right]
\end{aligned}$$

$$
\begin{aligned}
&+ (b-1)\left[\eta_j x_j^{b-2} + \eta_p x_p^{b-2}\right] \\
=& -(a-1)\left[\left(\boldsymbol{\kappa}_{,j}^\top \boldsymbol{x}^a\right) x_j^{a-2} + \left(\boldsymbol{\kappa}_{,p}^\top \boldsymbol{x}^a\right) x_p^{a-2}\right] \\
&- a\left[\kappa_{jj} x_j^{2a-2} + \kappa_{pp} x_p^{2a-2} + 2\kappa_{jp} x_j^{a-1} x_p^{a-1}\right] \\
&+ (b-1)\left[\eta_{0,j} x_j^{b-2} + \eta_{0,p} x_p^{b-2}\right] \\
&+ (b-1)\left[y_j \eta_{1,j} x_j^{b-2} + y_p \eta_{1,p} x_p^{b-2}\right]
\end{aligned}
$$

Plugging these definitions into the loss function in Eq. (6) and rearranging the individual terms in the same way as in Yu et al. (2024) yields $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ as follows (Figure 1c):

$$
\boldsymbol{\Gamma} \equiv \begin{bmatrix} \boldsymbol{\Gamma}_{\mathbf{K}} & \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta_0}} & \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta_1}} \\ \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta_0}}^\top & \boldsymbol{\Gamma}_{\boldsymbol{\eta_0}} & \boldsymbol{\Gamma}_{\boldsymbol{\eta_0},\boldsymbol{\eta_1}} \\ \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta_1}}^\top & \boldsymbol{\Gamma}_{\boldsymbol{\eta_0},\boldsymbol{\eta_1}}^\top & \boldsymbol{\Gamma}_{\boldsymbol{\eta_1}} \end{bmatrix} \in \mathbb{R}^{(p^2+2p)\times(p^2+2p)}, \quad \boldsymbol{g} \equiv \left(\mathrm{vec}(\boldsymbol{g}_{\mathbf{K}}), \boldsymbol{g}_{\boldsymbol{\eta_0}}, \boldsymbol{g}_{\boldsymbol{\eta_1}}\right) \in \mathbb{R}^{p^2+2p}, \tag{11}
$$

where $\boldsymbol{\Gamma}$ and $\mathbf{g}$ have a block structure with $\boldsymbol{\Gamma}_{\mathbf{K}} \in \mathbb{R}^{p^2\times p^2}$, $\boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta_0}} \in \mathbb{R}^{p^2\times p}$, $\boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta_1}} \in \mathbb{R}^{p^2\times p}$, $\boldsymbol{\Gamma}_{\boldsymbol{\eta_0}} \in \mathbb{R}^{p\times p}$, $\boldsymbol{\Gamma}_{\boldsymbol{\eta_0},\boldsymbol{\eta_1}} \in \mathbb{R}^{p\times p}$, $\boldsymbol{\Gamma}_{\boldsymbol{\eta_1}} \in \mathbb{R}^{p\times p}$, and $\mathbf{g}_{\mathbf{K}} \in \mathbb{R}^{p^2}$, $\boldsymbol{g}_{\boldsymbol{\eta_0}} \in \mathbb{R}^{p}$, $\boldsymbol{g}_{\boldsymbol{\eta_1}} \in \mathbb{R}^{p}$.

The exact derivations are shown in Appendix B. By recognizing that each entry of $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ can be written as a mean over all samples, the elements related to $\boldsymbol{\eta}_1$ can be computed directly from the elements related to $\boldsymbol{\eta}_0$:

$$
\begin{aligned}
\boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta_1}} &= \frac{1}{n}\sum_{i=1}^{n} y \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta_0}}^{(i)} \\
\boldsymbol{\Gamma}_{\boldsymbol{\eta_0},\boldsymbol{\eta_1}} &= \frac{1}{n}\sum_{i=1}^{n} y \boldsymbol{\Gamma}_{\boldsymbol{\eta_0}}^{(i)} \\
\boldsymbol{\Gamma}_{\boldsymbol{\eta_1}} &= \frac{1}{n}\sum_{i=1}^{n} y^2 \boldsymbol{\Gamma}_{\boldsymbol{\eta_0}}^{(i)} \\
\boldsymbol{g}_{\boldsymbol{\eta_1}} &= \frac{1}{n}\sum_{i=1}^{n} y \boldsymbol{g}_{\boldsymbol{\eta_0}}^{(i)}.
\end{aligned}
$$

Therefore, the computational overhead for computing the additional sub-matrices and sub-vectors related to $\boldsymbol{\eta_1}$ is negligible. Still, the addition of p dimensions to the optimization problem (Eq. 8) increases the problem dimensionality from $p^2 + p$ to $p^2 + 2p$ compared to the covariate-less model.

2.2.2 Model Identifiability through Regularization

Since the number of parameters in the power interaction model scales quadratic with p, real HTS data applications are in the high-dimensional regime with more parameters than samples, i.e., $p^2 + 2p > n$. To enable model identification, we place a ℓ_1 regularization penalty on the off-diagonal elements $\mathbf{K}_{\text{off}}$ of $\mathbf{K}$:

$$\hat{L}_{\boldsymbol{h},\boldsymbol{C},\lambda_1,\delta}(P_{\boldsymbol{\theta}}) = \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Gamma}_{\boldsymbol{\delta}}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top \boldsymbol{\theta} + \lambda_1 ||\text{vec}(\mathbf{K}_{\text{off}})||_1 \,. \tag{12}$$

As defined in section 2.2.1, $\boldsymbol{\theta} = (\text{vec}(\boldsymbol{K}), \boldsymbol{\eta_0}, \boldsymbol{\eta_1})$ comprises all model parameters, and $P_{\boldsymbol{\theta}}$ denotes the power interaction model with probability density $p_{\boldsymbol{\theta}}$ defined in Eq. 4. Following Yu et al. (2024), we multiply the diagonal entries of $\boldsymbol{\Gamma}(\boldsymbol{x})$ corresponding to $\boldsymbol{K}$ by a factor $\delta > 1$ to avoid an unbounded loss function. We denote $\boldsymbol{\Gamma}(\boldsymbol{x})$ with scaled diagonal entries as $\boldsymbol{\Gamma}_{\boldsymbol{\delta}}(\boldsymbol{x})$. Here, we use the default value from the implementation of Yu et al. (2024), $\delta = 2 - \frac{1}{1+4e\max(6\log(p)/n, \sqrt{6\log(p)/n})}$. In cases where $p \gg n$, the entries of $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ can be penalized as well with a regularization parameter λ_2:

$$\hat{L}_{\boldsymbol{h},\boldsymbol{C},\lambda_1,\lambda_2,\delta}(P_{\boldsymbol{\theta}}) = \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Gamma}_{\boldsymbol{\delta}}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top \boldsymbol{\theta} + \lambda_1 ||\text{vec}(\mathbf{K}_{\text{off}})||_1 + \lambda_2 ||\boldsymbol{\eta}_0||_1 + \lambda_2 ||\boldsymbol{\eta}_1||_1 \tag{13}$$

Furthermore, assuming $\mathbf{K}$ to be sparse matches the widely popular view of sparse association networks between microbial features or cell types (see, e.g., (Kurtz et al., 2015)). In the following, we will focus our attention on models without regularization on the location parameter.

Algorithm 1

Input: Initial estimate $\hat{\theta}^{(0)}$
Input: t_{max}, maximum number of iterations
Input: ϵ, the maximal tolerance level

Initialize $t \leftarrow 1$
Initialize $C \leftarrow \epsilon + 1$ (C stands for convergence criteria)
while $C > \epsilon$ or $t < t_{max}$ **do**
  $\hat{\theta}^{(t)} \leftarrow \hat{\theta}^{(t-1)}$
  for $j \leftarrow 1, 2, \ldots, s$ **do**
    $\hat{\theta}_j^{(t)} \leftarrow \text{Soft}\left(\frac{-(\boldsymbol{\Gamma}_{\boldsymbol{\delta}}(\mathbf{x})_{-j,j})^T \hat{\theta}_{-j}^{(t)} - g(\mathbf{x})_j}{\boldsymbol{\Gamma}_{\boldsymbol{\delta}}(\mathbf{x})_{jj}}, \frac{\lambda}{\boldsymbol{\Gamma}_{\boldsymbol{\delta}}(\mathbf{x})_{jj}}\right)$.
  end for
  $C \leftarrow \|\hat{\theta}^{(t)} - \hat{\theta}^{(t-1)}\|_1$
  $t \leftarrow t + 1$
end while

2.2.3 Computational implementation

The regularized score matching loss $\hat{L}_{\boldsymbol{h},\boldsymbol{C},\lambda,\delta}(P_{\boldsymbol{\theta}})$ (Eq. 12) represents a (large-scale) ℓ_1-penalized quadratic optimization problem that can be numerically solved with

a variety of optimization methods. Here, we follow Yu et al. (2024) and employ a proximal coordinate descent scheme (see also Algorithm 2 in Lin et al. (2016)). This algorithm also covers the covariate-extended a-b power interaction model and is described in Algorithm 1. Here, s is the dimensionality of $\boldsymbol{\theta}$ and Soft$(\cdot)$ is the softmax function. The default settings in `cosmoDA` are $\epsilon = 10^{-1}$ and $t_{max} = 1000$.

At its core, our Python implementation of Algorithm 1 uses the C implementation from the *genscore* package Yu et al. (2019, 2024) and included in the `cosmoDA` Python package. The `cosmoDA` package also provides an interface for a-b power interaction models that is equivalent to the R interface in the *genscore* package.

2.3 Differential Abundance Testing

One of the key objectives of `cosmoDA` is to determine the statistical significance of the covariate effects $\boldsymbol{\eta}_{1,j}$ for every feature $j = 1 \dots p$. Here, we combine results from Zhou et al. (2022) and Scealy and Wood (2022) to test the null hypothesis

$$\mathrm{H}_0 : \boldsymbol{\eta}_{1,j} = 0 \quad \text{against the alternative} \quad \mathrm{H}_1 : \boldsymbol{\eta}_{1,j} \neq 0 \,.$$

Let $\hat{\boldsymbol{\theta}} = (\mathrm{vec}(\hat{\boldsymbol{K}}), \hat{\boldsymbol{\eta}}_0, \hat{\boldsymbol{\eta}}_1)$ be the parameter estimates obtained from the score matching estimation framework. Scealy and Wood (2022) show that, under certain technical conditions and assumptions (see Theorem 1 and 2 in (Scealy and Wood, 2022)) the quantity $\hat{\mathbf{S}} = \boldsymbol{\Gamma}^{-1}(\boldsymbol{x})\hat{\boldsymbol{\Sigma}}_0\boldsymbol{\Gamma}^{-1}(\boldsymbol{x})$ yields a consistent estimator for $\mathrm{Var}(\hat{\boldsymbol{\theta}})$. In `cosmoDA`, we estimate $\hat{\boldsymbol{\Sigma}}_0$ as follows:

$$\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{n}\sum_{i=1}^{n}(\tilde{\boldsymbol{\Gamma}}_{\boldsymbol{\delta}}^{(i)}(\boldsymbol{x})\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{g}}^{(i)}(\boldsymbol{x}))(\tilde{\boldsymbol{\Gamma}}_{\boldsymbol{\delta}}^{(i)}(\boldsymbol{x})\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{g}}^{(i)}(\boldsymbol{x}))^T \,, \tag{14}$$

where $\tilde{\boldsymbol{\Gamma}}_{\boldsymbol{\delta}}^{(i)}(\boldsymbol{x})$ and $\tilde{\boldsymbol{g}}^{(i)}(\boldsymbol{x})$ are the components of $\tilde{\boldsymbol{\Gamma}}_{\boldsymbol{\delta}}(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{x})$ corresponding to the i-th sample. By selecting the components of $\hat{\mathbf{S}}$ corresponding to $\eta_{1,j}$, we derive the studentized test statistic

$$T_j = \hat{\eta}_{1,j}/\hat{S}_{\eta_{1,j}} \,, \tag{15}$$

which approximately follows a t-distribution with $n - 3$ degrees of freedom. The corresponding asymptotic p-values read:

$$p_j = 2F_{t,n-3}(-|T_j|) \,, \tag{16}$$

where $F_{t,n-3}$ is the cumulative distribution function of the t-distribution with $n-3$ degrees of freedom (Zhou et al., 2022). In `cosmoDA`, raw p-values are adjusted for multiple testing using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) (see Figure 1d for illustration).

2.4 Model selection and hyperparameter tuning

To make `cosmoDA` fully data-adaptive, we provide several strategies to select the hyperparameters of the framework. We first describe regularization parameter selection, followed by a novel data-driven approach to select the exponents a and b in a-b power interaction models.

2.4.1 Regularization parameter selection

`cosmoDA` provides several model selection methods to determine the regularization parameter λ_1 (and λ_2, respectively). The default strategy is k-fold cross-validation ($k = 5$) with the 1SE rule (Hastie et al., 2009). Here, the largest λ_1 value is chosen that lies within one standard error band of the λ_1 that minimizes the cross-validated regularized score matching loss (Eq. 12). The range of the λ_1-path is chosen to cover the whole range of possible sparsity of $\boldsymbol{K}$, i.e., from a fully dense $\boldsymbol{K}$ to a diagonal $\boldsymbol{K}$ (see Fig. E16b). This range depends on the dimensions of the dataset at hand and the chosen power transformation (see Appendix C). Per default, our implementation considers 100 λ_1-values log-linearly spaced in the interval $\left[10^{-6}, 1\right]$.

`cosmoDA` also allows λ_1 to be selected via the extended Bayesian Information Criterion (eBIC, Foygel and Drton (2010)). Following Yu et al. (2019), the eBIC for the a-b power interaction model reads:

$$\text{eBIC}_\gamma(\text{vec}(\mathbf{K}_{\text{off}})) = \text{S}(\text{vec}(\mathbf{K}_{\text{off}}))\log(n) - 2\log(\hat{L}_{\boldsymbol{h},\boldsymbol{C},\lambda_1,\delta}(P_{\boldsymbol{\theta}}) + 2\gamma||\text{vec}(\mathbf{K}_{\text{off}})||_1\,, \quad (17)$$

where $\text{S}(\text{vec}(\mathbf{K}_{\text{off}}))$ denotes the size of the support of $\text{vec}(\mathbf{K}_{\text{off}})$. The default γ value is $\gamma = 0.5$.

2.4.2 Data-driven selection of a-b powers

A key strength of a-b power interaction models is their seamless applicability to compositional data with excess zeros. While the limiting case $a = b = 0$ (i.e., log-transforming the data) requires a strategy for zero replacement or zero imputation (Lubbe et al., 2021; Greenacre et al., 2023) with potentially detrimental effects for downstream analysis (Te Beest et al., 2021), we propose a data-driven tuning strategy for power interaction models with powers $a > 0$ and $b > 0$ that keeps the original data unaltered. For simplicity, we consider the setting $a = b$.

We first note that the power transformations in the models 2 and 4 are similar to the Box-Cox transformation (Box and Cox, 1964) of the form:

$$x_\phi = \begin{cases} \frac{1}{\phi}(x^\phi - 1), & \text{if } \phi > 0 \\ \log x, & \text{if } \phi = 0\,, \end{cases} \quad (18)$$

with $\lim_{\phi\to 0}\frac{1}{\phi}(x^\phi - 1) = \log(x)$ (see also Fig. 1b for illustration). The Box-Cox transformation and the power transformation used in a-b power interaction models (Eqs. 2, 4) are, however, not equivalent due to the -1 term in the Box-Cox transformation. By introducing scaling factors for the score matching elements in Eq. 11, we can nevertheless achieve the same asymptotic approximation to the logarithm as the Box-Cox transformation (see Appendix C for details).

While ϕ is typically tuned to make the transformed data approach normality, we follow a geometric strategy inspired by the one presented in Greenacre (2024). Specifically, we determine $\phi = a = b$ to let the resulting "geometry" of transformed data be

as similar as possible to the appropriate log-ratio geometry. This is achieved by maximally aligning the principal component (PC) embedding of log-ratio transformed data with imputed zeros and the PC embedding of a power transformation with parameter ϕ of the data with zero entries Tsagris et al. (2016). Maximal alignment is defined as the highest Procrustes correlation of both embeddings over a range of values $\phi \in]0, 1[$. In the case of power interaction models, it is natural to select a power that closely matches the geometry of data after the additive log-ratio (ALR) transform, since a-b power interaction models with $\phi = a = b = 0$ are a generalization of Aitchison's A^{p-1} distributions after ALR transformation of the data (Aitchison and Shen, 1980). Since equal dimensionality of the ALR and power-transformed data is required, we append the column $\log(\frac{\mathbf{X}_p}{\mathbf{X}_p}) = \mathbf{0}_p$ to the ALR transformation of $\mathbf{X}$ before performing PC analysis.

The original procedure to obtain maximal Procrustes correlation is outlined in Greenacre (2024). We use the same procedure, but with different input matrices. Let $\boldsymbol{X}_\phi$ be the Box-Cox-like transformed data with

$$\boldsymbol{X}_{\phi,j} = \frac{1}{\phi}(p\frac{\boldsymbol{X}_j^\phi}{\sum_{k=1}^p \boldsymbol{X}_k^\phi} - 1)\,, \tag{19}$$

and $\boldsymbol{X}_{ALR}$ is the ALR-transformed data (with pseudocount 0.5 for all zeros) with column $\mathbf{0}_p$ appended.

We compute the Procrustes correlation r_ϕ between the two data matrices as follows:

(i)	Matrix normalization:	$\boldsymbol{X}^*_\phi$	$= \boldsymbol{X}_\phi / \sqrt{\text{trace}(\boldsymbol{X}_\phi^T \boldsymbol{X}_\phi)}$
		$\boldsymbol{X}^*_{ALR}$	$= \boldsymbol{X}_{ALR} / \sqrt{\text{trace}(\boldsymbol{X}_{ALR}^T \boldsymbol{X}_{ALR})}$
(ii)	SVD of cross product:	$\boldsymbol{S}$	$= \boldsymbol{X}_\phi^{*T} \boldsymbol{X}^*_{ALR} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$
(iii)	Optimal rotation matrix:	$\boldsymbol{Q}$	$= \boldsymbol{V}\boldsymbol{U}^T$
(iv)	Sum of squared errors:	E_ϕ	$= \text{trace}((\boldsymbol{X}^*_\phi - \boldsymbol{X}^*_{ALR}\boldsymbol{Q})^T(\boldsymbol{X}^*_\phi - \boldsymbol{X}^*_{ALR}\boldsymbol{Q}))$
(v)	Procrustes correlation:	r_ϕ	$= \sqrt{1 - E_\phi}$

For a given dataset, the optimal power ϕ^* is determined by $\phi^* := \arg\max_\phi r_\phi$ for $\phi \in]0, 1[$.

3 Simulation benchmarks

We next provide two simulation studies that benchmark two key features of `cosmoDA`: (i) sparse recovery of feature interactions in the covariate-extended a-b power interaction model and (ii) identification of differentially abundant features in the presence of feature correlations. The first benchmark complements the extensive covariate-free simulation benchmarks of Yu et al. (2024), the second one provides a new realistic semi-synthetic simulation and evaluation setup, incorporating scRNA-Seq data Perez et al. (2022).

3.1 Sparse recovery of feature interactions in the presence of a covariate

One of the core strengths of a-b power interaction models is their ability to recover (potentially) sparse feature interaction matrices $\mathbf{K}$. Yu et al. (2024) provide an extensive simulation framework that evaluates the influence of hyperparameters, sample size, and interaction topologies on recovery performance of the a-b power interaction model. We focus here on evaluating the influence of covariate inclusion on the model's ability to identify sparse feature interactions. Specifically, we expect interaction recovery to be *independent of covariate inclusion.*

Following Yu et al. (2024), we generated compositional data $\boldsymbol{X} \in \Delta_{p-1}^{n}$ from an A^{p-1} model with $p = 100$ features using the model in Eq. 4 with the constraint that $\boldsymbol{x}^T \boldsymbol{K} \boldsymbol{x} > 0 \; \forall \boldsymbol{x}, \boldsymbol{\eta} \succeq -1$. To probe sample size dependencies, we used two scenarios $n = 80$ and $n = 1000$, respectively. We set $\boldsymbol{\eta}_0 = -\mathbf{1}_p$, and considered banded interaction matrices $\mathbf{K}$ with bandwidths $s = 2$ if $n = 80$ and $s = 7$ if $n = 1000$, as suggested by Yu et al. (2024). We further defined the nonzero off-diagonal entries of $\mathbf{K}$ as $\mathbf{K}_{i,j} = |i - j|/(s + 1) - 1$ for all $i \neq j, 1 \leq |i - j| \leq s$, and the diagonal entries as the negative sum of the off- diagonals, to ensure the sum-to-zero constraint on the rows of $\mathbf{K}$ (Figure E2). This definition slightly deviates from the definition in Yu et al. (2024), as the sign of all entries in $\mathbf{K}$ is flipped, but ensures positive definiteness of $\mathbf{K}$. This modification allows the efficient use of the adaptive rejection sampler for data generation, as provided in the *genscore* R package (Yu et al., 2019). For both sample sizes, we generated $R = 50$ replicates of the data.

We applied three different methods for regularized estimation of the underlying interaction matrix $\mathbf{K}$ to all datasets:

1. The a-b power interaction model ($a = b = 0$) without covariate (Eq. 2). This model allows the estimation of $\mathbf{K}$ and $\boldsymbol{\eta_0}$. We estimated these models through the implementation in `cosmoDA`.
2. The covariate-extended a-b power interaction model ($a = b = 0$) (Eq. 4). Here, we used a misspecified $\boldsymbol{y}$ where each entry is drawn uniformly at random from $\{0, 1\}$. The model allows the estimation of $\mathbf{K}, \boldsymbol{\eta}_0$, and $\boldsymbol{\eta}_1$. We used the implementation in `cosmoDA`.
3. The graphical lasso model on CLR-transformed data, as introduced in *SPIEC-EASI* (Kurtz et al., 2015). The non-zero entries of the resulting sparse inverse

covariance matrix serve as a (mis-specified) proxy for $\mathbf{K}$. We used the implementation from the *gglasso* package (Schaipp et al., 2021). Model selection was performed with the extended BIC (eBIC) criterion (Foygel and Drton, 2010) with $\gamma = 0.25$.

For all three models, we used $n_\lambda = 100$ values for the regularization parameter, log-spaced in the range $10^{-6} < \lambda_1 < 1$, and $k = 5$ cross-validation folds. All score matching estimation parameters were set to the defaults recommended by Yu et al. (2024) (see also Section 2.2.3).

To measure recovery performance, we compared the support of the off-diagonal elements of the estimated $\hat{\mathbf{K}}$ and the ground truth $\mathbf{K}$ by calculating the true positive rate (TPR) and true negative rate (TNR), and assessing them through Receiver operating characteristic (ROC) curves.

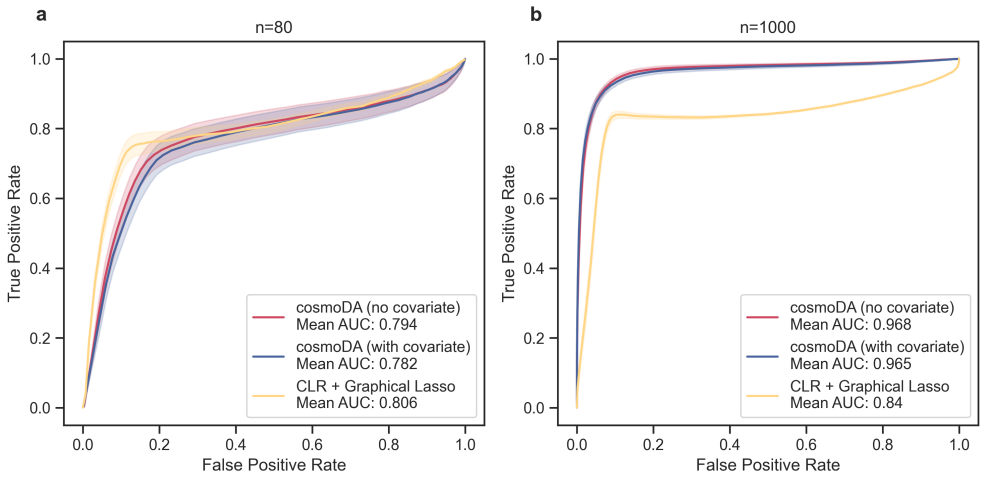


Fig. 2 Recovery of K improves with sample size and is not impacted by covariate inclusion. Receiver operating curves for `cosmoDA` with and without covariate effect estimation, as well as CLR transform and graphical lasso for with (a) $n = 80$ and (b) $n = 1000$. The solid lines depict the mean ROC over all 50 generated datasets, the shaded areas show the standard error.

Figure 2 summarizes the average ROC curves for the two different sample sizes. For $n = 80$ (Figure 2a), we observed that both a-b power interaction models showed almost equivalent ability to reconstruct the interaction matrix (mean AUC 0.782 vs. 0.794). Their performance was slightly worse that the graphical lasso (mean AUC 0.806), especially for false positive rates smaller than 0.2. When increasing the sample size to $n = 1000$, all three methods showed improvements in recovery performance, improving the mean average AUC as well as reducing the variance in results (Figure 2b). As expected, including a covariate in the a-b power interaction model had only marginal impact on the mean AUC (0.965 vs. 0.968). Contrary to the low sample size case, both a-b power interaction models significantly outperformed the (misspecified) graphical lasso (mean AUC 0.84) across the entire range of regularization strengths.

3.2 Differential abundance testing in the presence of correlated features

To test the effectiveness of `cosmoDA` in detecting differentially abundant features in the presence of realistic feature interactions, we designed the following semi-synthetic simulation benchmark.

We considered a scRNA-seq data set from Perez et al. (2022) that derived relative abundance values of $p = 11$ types of peripheral blood mononuclear cells (PBMCs) from overall $n = 352$ samples. The samples come from 260 unique subjects, 162 of which are patients with with systemic lupus erythematosus (SLE) (208 samples) and 98 healthy controls (144 samples). We used these data to estimate realistic base values for the interaction matrix $\mathbf{K}$ and the location vector $\boldsymbol{\eta}$, respectively. The base model is the a-b power interaction model without covariate (Eq. 2). We set $a = b = 0$ and used $\lambda_1 = 0.043$ as sparsity parameter. We considered the NK cell type as the pth reference component for all power interaction models due to their high abundance and low variance between groups. The resulting interaction matrix $\mathbf{K}_B$ and location vector $\boldsymbol{\eta}_{0,B}$ are shown in Figure 3.

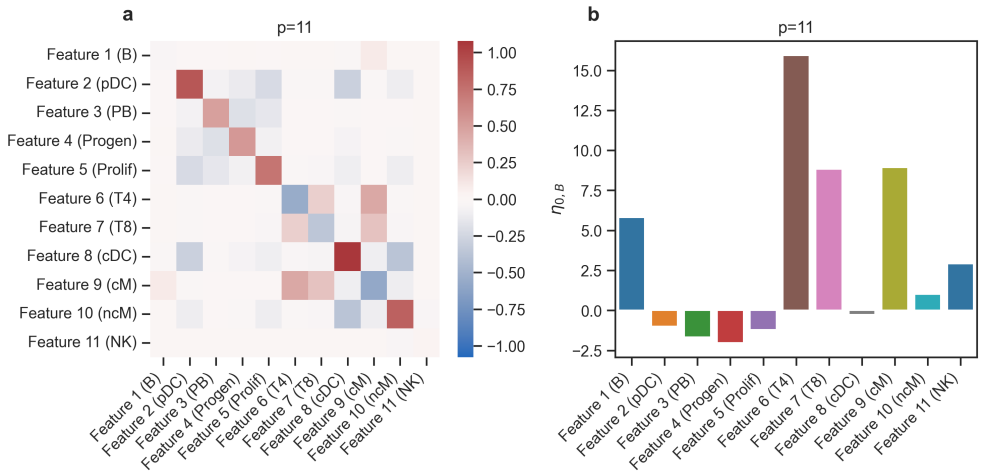


Fig. 3 Data generation parameters used for the Differential abundance testing benchmark, $p = 11$. Parameters were generated by running the power interaction model without covariate on the dataset from Perez et al. (2022). The names of the cell types from the original dataset are shown in brackets. (a) Interaction matrix ($\boldsymbol{K}_B$). (b) Location vector ($\boldsymbol{\eta}_{0,B}$).

To generate ground-truth differentially abundant cell types, we defined the effect vector $\boldsymbol{\eta}_{1,B} = \tau \boldsymbol{i} \boldsymbol{\eta}_{0,B}$, where $\boldsymbol{i}$ is a p-dimensional binary vector that indicates the cell types that are influenced by the condition (i.e., are differentially abundant), and $\tau = (-0.5, -0.3, 0.3, 0.5, 1)$ controls the relative effect size.

Using this model, we considered three differential abundance scenarios: (i) Estimation when the effect is on a rare cell type with (pDC), (ii) effect estimation on an

abundant cell type (T4), and (iii) effects on both cell types (pDC and T4). Two different sample sizes ($n = 100$ and $n = 1000$) were considered for each case. For each of the resulting 30 scenarios, we generated five datasets with $n/2$ control samples ($\mathbf{K} = \mathbf{K}_B$, $\boldsymbol{\eta} = \boldsymbol{\eta}_{0,B}$) and $n/2$ case samples ($\mathbf{K} = \mathbf{K}_B$, $\boldsymbol{\eta} = \boldsymbol{\eta}_{0,B} + \boldsymbol{\eta}_{1,B}$). To simulate these semi-synthetic data sets, we used the adaptive rejection sampler from the *genscore* R package (Yu et al., 2019).

To showcase the performance of `cosmoDA` for higher dimensional datasets, we conducted another set of simulations with $p = 99$ features. We constructed the corresponding interaction matrix as a block-diagonal matrix, using the original $\mathbf{K}_B$ matrix in each of the nine blocks (see Figure E4a). Likewise, we stacked the scenario-specific location vectors $\boldsymbol{\eta}_{0,B}$ and $\boldsymbol{\eta}_{1,B}$ nine times to obtain the high-dimensional location vectors (Figure E4b).

We compared the ability of four different DA testing methods to recover differentially abundant features at an expected FDR level of $\alpha = 0.05$:

- DA testing with `cosmoDA` ($a = b = 0$). We used $n_\lambda = 20$ values between 10^{-3} and 2 for λ_1 and 5-fold cross validation with the 1SE rule for model selection. All other parameters were set to default values described in Section 2.2.3).
- ANCOM-BC (Lin and Peddada, 2020) with default parameters as an example for a common DA testing method without feature interactions. Since ANCOM-BC assumes count data instead of relative abundances, we scaled the simulated data by the median sequencing depth over all samples in the original dataset and rounded to the nearest integer to obtain comparable counts.
- A Dirichlet regression model and subsequent significance test on the regression coefficients, as implemented by Maier (2014). This model serves as a simple baseline that does not respect feature interactions.
- CompDA (Ma et al., 2024), a recent DA testing method for compositional data, respecting feature interactions via conditional dependency modeling.

Figure 4 summarizes the results for the simulation scenario with the original number of features ($p = 11$). Here, `cosmoDA` showed the overall best ability to recover the true effects (Matthews' correlation coefficient, Figure 4a), especially when the sample size was larger and for the more abundant cell type T4 (see Figure E5). Importantly, `cosmoDA` showed the lowest FDR value in all scenarios (Figure 4b). Although `cosmoDA` was not able to control the FDR at the expected level of 0.05 in every scenario, the methods without consideration of interactions (ANCOM-BC and Dirichlet) detected more false positive features, with FDR levels averaging between 0.2 and 0.7 in most scenarios. Surprisingly, CompDA did not achieve lower FDR values than ANCOM-BC and performed worse than Dirichlet regression in all cases. We observed slightly elevated FDR levels of `cosmoDA` in cases where the DA cell types were not detected, resulting in FDR values of 1 where one feature was falsely discovered (see Figure E6). While Dirichlet regression and ANCOM-BC struggled with FDR control in all scenarios (see Figure E6), CompDA produced much higher FDR values for the abundant cell type (T4). For smaller sample sizes ($n = 100$) and small effects, `cosmoDA` was not able to consistently detect the differentially abundant features, resulting in lower power for

these scenarios. With increasing sample size, the power of `cosmoDA` was on par with the other methods (Figures 4c, E7).

In the large-dimensional case ($p = 99$), the performance of all methods decreased in the small sample size scenario, while Matthews' correlation coefficient was similar to the case of $p = 11$ for larger sample sizes (Figures 5a, E11). Again, `cosmoDA` always showed the lowest FDR, albeit with slightly elevated levels for $n = 1000$, and mean FDR levels between 0.1 and 0.4 for $n = 100$ (Figure 5b). The FDR levels for `cosmoDA` did not show a trend across feature rarity and effect size, while the other methods were not able to produce average FDR levels below 0.5 for effects on rare pDC cells (Figure E12). In terms of power, only ANCOM-BC and Dirichlet regression were able to correctly detect some differentially abundant features for $n = 100$, while all methods showed good power for larger sample sizes (Figure 5c). Breaking these results down by cell type revealed a good power of ANCOM-BC and Dirichlet regression for abundant features, while effects on rare features could not be reliably detected by any method (Figure E13). The unsuitability of `cosmoDA` and CompDA for the case of $p = 99, n = 100$ is not surprising, as both models need to estimate pairwise feature interactions in the high-dimensional regime.

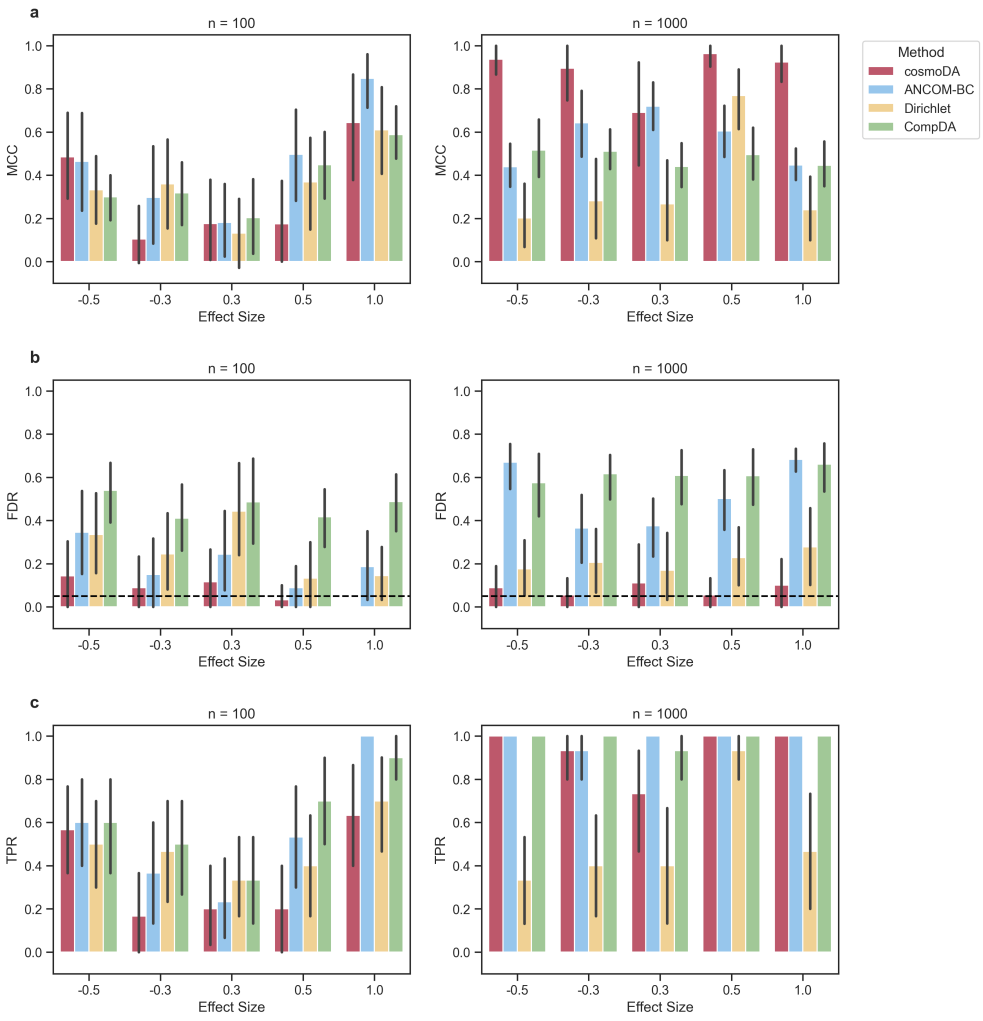


Fig. 4 Performance comparison for recovering differentially abundant features across different scenarios, $p = 11$**.** (a) Matthews' correlation coefficient. (b) False discovery rate. The dashed line shows the nominal FDR for all methods. (c) True positive rate (power).

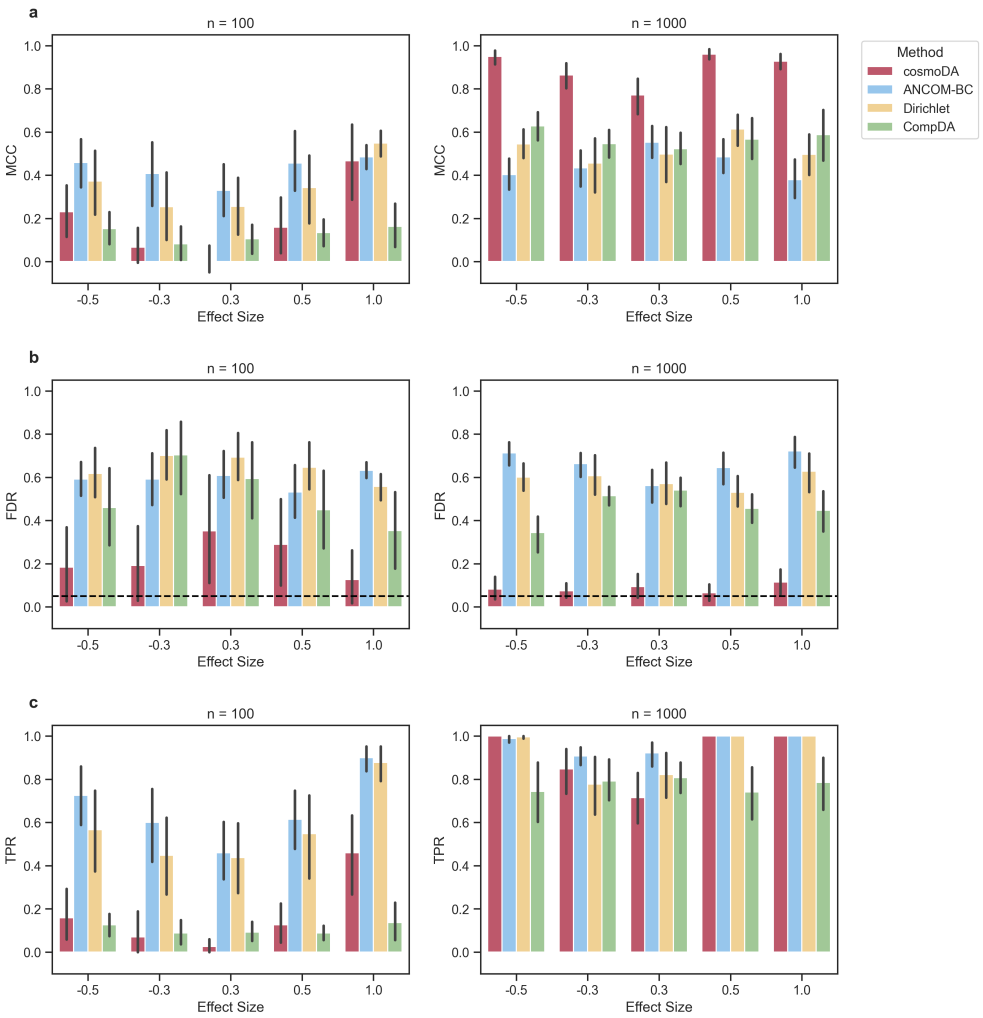


Fig. 5 Performance comparison for recovering differentially abundant features across different scenarios, $p = 99$**.** (a) Matthews' correlation coefficient. (b) False discovery rate. The dashed line shows the nominal FDR for all methods. (c) True positive rate (power).

4 Applications to single-cell and microbiome data

To showcase the DA testing capabilities of `cosmoDA` on real data, we considered two compositional datasets: PBMC abundances derived from scRNA-seq data of SLE patients (as used in the semi-synthetic benchmarks Perez et al. (2022)) and infant gut microbiome data from 16S rRNA sequencing (Yatsunenko et al., 2012). Apart from comparing the empirical results with other state-of-the-art methods, we also evaluated the impact of power transformations ($a = b = \phi \neq 0$) on the downstream DA results.

4.1 DA analysis of cell type compositions in patients with systemic lupus erythematosus

We used the scRNA-Seq-derived PBMC data from Perez et al. (2022) ($n = 352, p = 11$, see Section 3.2) to estimate differences in cell type composition between subjects with systemic lupus erythematosus (SLE) (n=208) and healthy controls (n=144).

To tune the parameters in `cosmoDA`, we considered the power values $\phi = (0.01, 0.02, 0.03, \ldots 0.99)$, as well as the log-log model ($\phi = 0$) for comparison. We set the range of λ_1 values between 1.5 and 10^{-7} to ensure full coverage of the range of supports of $\boldsymbol{K}$ for every value of ϕ. As before, we used NK cells as the reference cell type for `cosmoDA` and selected the regularization strength via 5-fold cross-validation with the 1SE rule. We used ANCOM-BC, Dirichlet regression, and CompDA for comparison.

We first investigated the influence of our power value tuning schemes for DA analysis. The Procrustes correlation analysis showed that the ALR-transformed PBMC data (with zeros replaced by a pseudocount of 0.5) and the power-transformed data had the highest alignment for a power of $\phi^* = 0.22$ (see Figure 6a). To investigate the impact of zero replacement and the power transform on differential abundance, we also compared the DA testing results of `cosmoDA` for all values of ϕ with and without zero entries (Figure 6c, d). Due to the low number of zeroes (4.5%), the impact of zero imputation was negligible for this dataset, making the adjusted p-values with and without zero imputation almost identical. Below a value of 0.8, the exponent of the power transformation only impacted the differential abundance of CD14+ classical monocytes (cM). For higher exponents, almost all cell types showed no differential abundance. Comparing the results of `cosmoDA` with ANCOM-BC, Dirichlet regression, and CompDA (all implemented as described in section 3.2) showed that all methods selected different sets of differentially abundant cell types (Figure 6b). CompDA produced the most conservative results, only finding four DA cell types at an FDR level of 0.05. On the other hand, Dirichlet regression found all cell types to be differentially abundant. Interestingly, `cosmoDA` was the only method that did *not* select classical monocytes as differentially abundant (Figure 6b). The latter finding is in agreement with a control experiment performed by Perez et al. (2022) that found absolute monocyte abundances to be not differentially abundant in SLE patients.

4.2 DA analysis in microbiome data

To showcase the suitability of `cosmoDA` for microbial 16S rRNA sequencing, we used gut microbiome data from infants in Malawi and the United States (Yatsunenko et al.,

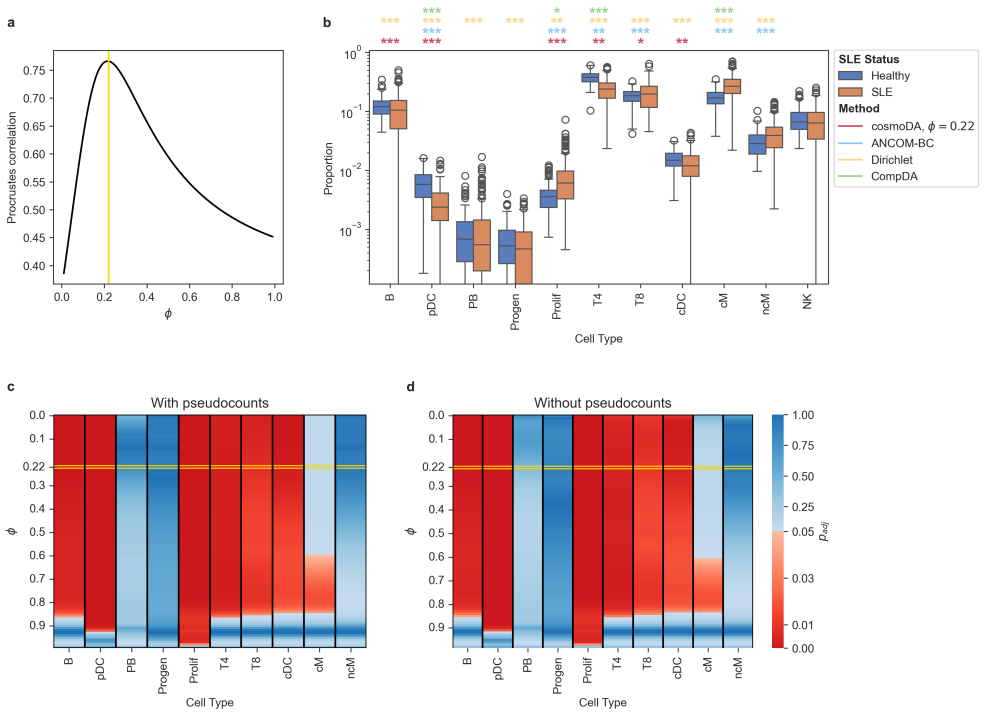


Fig. 6 Differential abundance testing with `cosmoDA` on the lupus dataset. (a) Procrustes correlation between power transformation and ALR transformation with zero replacement. The yellow line ($\phi^* = 0.22$) indicates the maximal Procrustes correlation. (b) Boxplot of relative abundance data without zero replacement. The colored stars indicate the significance level for each method (*: $p_{adj} < 0.05$; **: $p_{adj} < 0.01$; ***: $p_{adj} < 0.001$). Differential abundance results on NK cells (reference in cosmoDA) are omitted. (c) Adjusted p-values for testing differential abundance with `cosmoDA` on zero-replaced data with different power transformations. Red entries denote differential abundance at a level of $\alpha = 0.05$, blue entries denote no differential abundance. The yellow box highlights the adjusted p-values for ϕ^* determined in a. (d) Same as c, but using the raw data without zero replacement.

2012). We followed the pre-processing in the original ANCOM-BC study Lin and Peddada (2020) and aggregated the data to the Phylum level. We selected all samples from subjects aged less than two years old in Malawi and the United States. Following Lin and Peddada (2020), we next discarded all phyla where more than 90% of samples contained zero entries, resulting in $n = 97$ samples of $p = 13$ phyla. We selected Bacteroidetes as the reference phylum and applied `cosmoDA` with $\phi = (0.01, 0.02, 0.03, \ldots 0.99)$, and the log-log model ($\phi = 0$) to the relative abundances with and without zero replacement. The range of values for λ_1 was set to $\left[10^{-12},\ 1.5\right]$, and we used 5-fold cross-validation with the 1SE rule to select λ_1 for each value of ϕ.

For this dataset, our power selection scheme identified $\phi^* = 0.13$ to result in the best Procrustes alignment (see Figure 7a). The larger proportion of zero entries in this dataset (28.6%) caused more differences in downstream DA testing results, both on the original and zero-imputed data (Figure 7c, d). While the DA pattern of taxa with

no zero entries (Firmicutes, Actinobacteria, Tenericutes, and Proteobacteria) was not impacted by zero imputation, the phyla with at least 20% zero entries (Cyanobacteria, Elusimicrobia, Euryarchaeota, Lentispherae, Spirochaetes, and TM7) were deemed differentially abundant at a smaller power values. Similar to the analysis of the scRNA-seq data, the four DA methods produced different sets of DA taxa at an FDR level of 0.05 (Figure 7b). Dirichlet regression and CompDA seemed to be only sensitive to taxa with high average abundance, while ANCOM-BC and `cosmoDA` were able to also detect differential abundance in rare phyla. The set of DA taxa discovered by `cosmoDA` at $\phi^* = 0.13$ on the data with zero entries was smaller than the set discovered on the same dataset by ANCOM-BC. Nevertheless, `cosmoDA` found multiple phyla that are associated with rural lifestyles (Elusimicrobia, Euryarchaeota, Spirochaetes) to be increased in infants from Malawi (Herlemann et al., 2007; Obregon-Tito et al., 2015). Notably, the ANCOM-BC algorithm involves the replacement of zeros by a small pseudocount (Lin and Peddada, 2020). Indeed, the set of DA phyla discovered by `cosmoDA` on the zero-replaced data with the exponent $\phi^* = 0.13$ (Figure 7c) almost perfectly matched the DA phyla found by ANCOM-BC (except Firmicutes and Proteobacteria). Overall, this confirms that replacement of zero entries in microbial abundance data has significant impact on differential abundance.

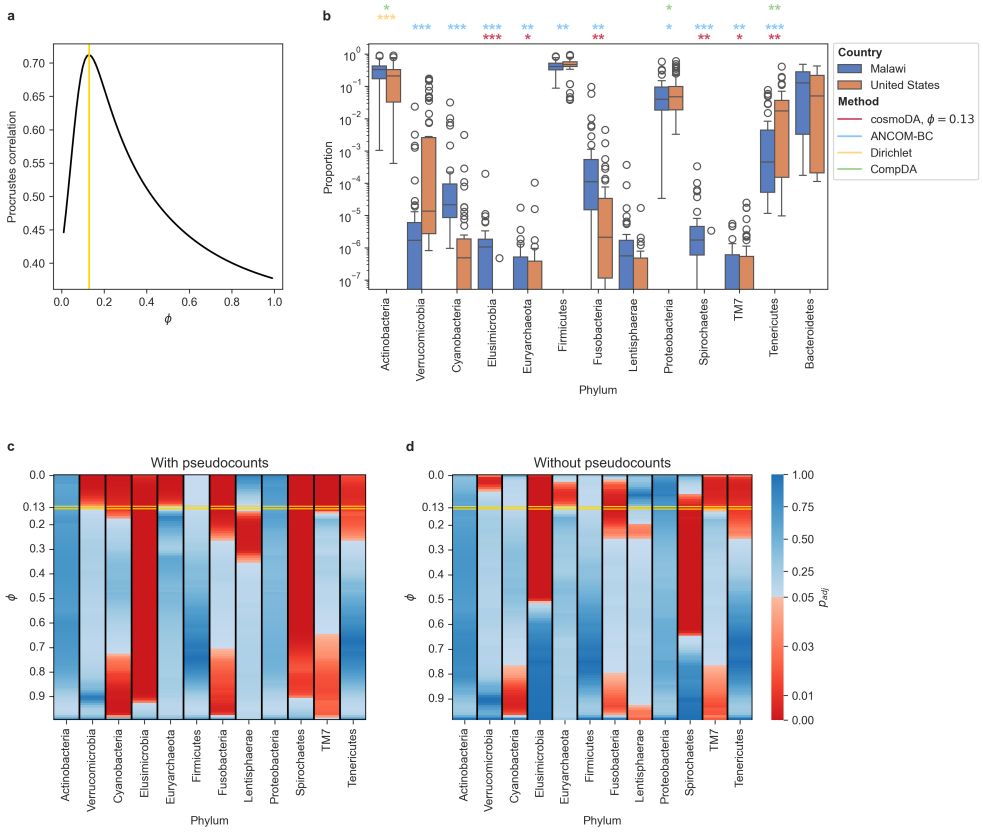


Fig. 7 Differential abundance testing (US vs. Malawi) with `cosmoDA` on infants (age $<$ 2 years) in the the human gut dataset. (a) Procrustes correlation between power transformation and ALR transformation with zero replacement. The yellow line ($\phi^* = 0.13$) indicates the maximal Procrustes correlation. (b) Boxplot of relative abundance data without zero replacement. The colored stars indicate the significance level for each method (*: $p_{adj} < 0.05$; **: $p_{adj} < 0.01$; ***: $p_{adj} < 0.001$). Differential abundance results on Bacteroidetes (reference in cosmoDA) are omitted. (c) Adjusted p-values for testing differential abundance with `cosmoDA` on zero-replaced data with different power transformations. Red entries denote differential abundance at a level of $\alpha = 0.05$, blue entries denote no differential abundance. The yellow box highlights the adjusted p-values for ϕ^* determined in a. (d) Same as c, but using the raw data without zero replacement.

5 Conclusion

Tissues and bacterial communities are complex biological environments, governed by interactions between individual cell types or microbial taxa. The prevailing high-throughput sequencing (HTS) data sets probing these complex mixtures are often compositional in nature. Statistical generative modeling as well as differential abundance testing schemes for such compositional datasets can therefore suffer from inaccuracies if interactions between cells or microbes are not considered in the analysis. Extending the class of a-b power interaction models (Yu et al., 2024) by a linear effect on the location vector, our new method `cosmoDA` allows to accurately model HTS data with pairwise feature interactions in the presence of covariate information. The covariate formulation in `cosmoDA` also seamlessly integrates into the generalized score matching optimization framework (Hyvärinen, 2005; Lin et al., 2016; Yu et al., 2022), facilitating fast and accurate parameter inference. L_1 regularization on the interaction matrix further avoids model complexity explosion and allows to select parsimonious interaction patterns. Compared to the a-b power interaction model without covariates from Yu et al. (2024), the addition of a covariate did not reduce its ability to detect significant feature interactions in our synthetic data simulations. Both the covariate-less and covariate-extended a-b power interaction models outperformed other established procedures for identifying sparse interactions in compositional HTS data when the sample size was sufficiently large.

In the presence of a binary condition, testing for significance of the covariate-related parameters in the location vector acts as a form of differential abundance testing. Here, the parallel estimation of feature interactions helps to avoid false positive detections which are only indirectly related to the condition. In our realistic simulation experiments, `cosmoDA` was the only method to approximately control the false discovery rate in the presence of feature interactions, while no other tested method could distinguish between direct and indirect compositional changes. `cosmoDA` showed reduced power when the sample size was small, but was on par with methods like ANCOM-BC (Lin and Peddada, 2020) for larger numbers of observations. One exception where `cosmoDA` was not able to adequately control the FDR was for misspecified models with more features than samples. We further demonstrated the ability of `cosmoDA` to find biologically meaningful differential abundances on two experimental datasets from human single-cell RNA sequencing and microbiome 16S rRNA sequencing.

The use of power transformations instead of the logarithm in a-b power interaction models allows to keep zero measurements in the data as-is, avoiding distortions caused by imputation of these values. Through a small adjustment in the score matching optimizer, we were able to approximate the log-transformation for exponents approaching zero. Applying `cosmoDA` to real-world single-cell and microbiome datasets, we discovered that zero replacement and the exponent of the power transformation had a considerable impact on downstream DA results in data with excess zeros (Gloor et al., 2017). We further demonstrated that selecting an exponent for the power transformation that approximates the data geometry after an ALR transformation generally produces sensible differential abundance results.

While `cosmoDA` successfully tackles multiple challenges in generative modeling and differential abundance testing, it also has some limitations. Currently, `cosmoDA` can

only accommodate a single binary or continuous covariate. Extending the linear model formulation would allow to model more complex scenarios and adjust for confounders in DA testing. For this, the score matching estimator would also have to be extended to multiple covariates. The implementation of such a flexible model could be simplified by using automatic differentiation for determining the elements of $\mathbf{\Gamma}$ and $\boldsymbol{g}$ (Kassel et al., 2024). In addition, we believe that approximation of the logarithm for small exponents can be solved more elegantly by changing the general definition of a-b power interaction models to utilize a true Box-Cox transformation rather than using our proposed adjustments in the score matching optimizer. Estimation of our model also relies on selecting a good reference, which is profiled out in the model formulation. Looping over multiple references and averaging the results, as described by Yu et al. (2024), could avoid this dependency at the cost of computational efficiency.

While we empirically showed the feasibility of `cosmoDA`, we did not provide any guarantees for goodness of fit and convergence. A formal reevaluation and extension of the theoretical considerations provided by Yu et al. (2024) would give more justification to our approach.

Overall, we believe that `cosmoDA` with its abilities to include feature interactions and seamless handling of excess zeros represents a valuable addition to the growing family of differential abundance testing methods. A Python implementation of `cosmoDA` and the power interaction model without covariates is available at https://github.com/bio-datascience/cosmoDA.

Acknowledgments

We thank Matthias Drton for his inputs to the score matching optimization in `cosmoDA`, and Fabian Schaipp for his recommendations regarding the optimization algorithm. We also thank Oleg Vlasovets for implementing the graphical lasso on the simulated benchmark data. H.L. was supported by NIH grant R01GM123056.

Author contributions

J.O. developed the idea of `cosmoDA` with help from H.L. and C.L.M. J.O. implemented the method, performed and evaluated all simulations and conducted the data applications. J.O. wrote the manuscript with assistance from C.L.M. All authors read and approved the final manuscript.

Declaration of Interests

The authors declare no competing interests.

Data and code availability

All datasets used in this article are publicly available. The SLE scRNA-seq data was downloaded from the the Human Cell Atlas platform (GSE174188). The gut microbiome data is stored at MG-RAST https://www.mg-rast.org/index.html under search

string "mgp401", code for data preparation was adapted from https://github.com/FrederickHuangLin/ANCOM-BC. Code for reproducing the analyses in this article is available under https://github.com/bio-datascience/cosmoDA, intermediate data can be found at https://zenodo.org/records/13911623.

References

Aitchison, J.: The statistical analysis of compositional data. J. R. Stat. Soc. Series B Stat. Methodol. **44**(2), 139–160 (1982)

Aitchison, J.: A general class of distributions on the simplex. J. R. Stat. Soc. Series B Stat. Methodol. **47**(1), 136–146 (1985) https://doi.org/10.1111/j.2517-6161.1985.tb01341.x

Aitchison, J., Shen, S.M.: Logistic-normal distributions: Some properties and uses. Biometrika **67**(2), 261–272 (1980) https://doi.org/10.2307/2335470

Box, G.E.P., Cox, D.R.: An analysis of transformations. J. R. Stat. Soc. Series B Stat. Methodol. **26**(2), 211–243 (1964) https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Billheimer, D., Guttorp, P., Fagan, W.: Statistical interpretation of species composition. Journal of the American Statistical Association **96**(456), 1205–1214 (2001) https://doi.org/10.1198/016214501753381850

Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. **57**(1), 289–300 (1995)

Büttner, M., Ostner, J., Müller, C.L., Theis, F.J., Schubert, B.: scCODA is a bayesian model for compositional single-cell data analysis. Nat. Commun. **12**(1), 6876 (2021) https://doi.org/10.1038/s41467-021-27150-6

Erb, I.: Partial correlations in compositional data analysis. Applied Computing and Geosciences **6**, 100026 (2020)

Foygel, R., Drton, M.: Extended bayesian information criteria for gaussian graphical models. Advances in neural information processing systems **23** (2010)

Greenacre, M., Grunsky, E., Bacon-Shone, J., Erb, I., Quinn, T.: Aitchison's compositional data analysis 40 years on: A reappraisal. SSO Schweiz. Monatsschr. Zahnheilkd. **38**(3), 386–410 (2023) https://doi.org/10.1214/22-STS880

Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., Egozcue, J.J.: Microbiome datasets are compositional: And this is not optional. Front. Microbiol. **8**, 2224 (2017) https://doi.org/10.3389/fmicb.2017.02224

Greenacre, M.: The chiPower transformation: a valid alternative to logratio transformations in compositional data analysis. Adv. Data Anal. Classif. (2024) https://doi.org/10.1007/s11634-024-00600-x

Herlemann, D.P.R., Geissinger, O., Brune, A.: The termite group I phylum is highly diverse and widespread in the environment. Appl. Environ. Microbiol. **73**(20), 6682–6685 (2007) https://doi.org/10.1128/AEM.00712-07

Hijazi, R.H., Jernigan, R.W.: Modelling compositional data using dirichlet regression models. Journal of Applied Probability & Statistics **4**(1), 77–91 (2009)

Heumos, L., Schaar, A.C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M.D., Strobl, D.C., Henao, J., Curion, F., Single-cell Best Practices Consortium, Schiller, H.B., Theis, F.J.: Best practices for single-cell analysis across modalities. Nat. Rev. Genet., 1–23 (2023) https://doi.org/10.1038/s41576-023-00586-w

Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, 2nd edn. Springer series in statistics. Springer, New York, NY (2009). https://doi.org/10.1007/978-0-387-84858-7 . https://link.springer.com/book/10.1007/978-0-387-84858-7

Hyvärinen, A.: Estimation of Non-Normalized statistical models by score matching. Journal of Machine Learning Research **6**, 695–709 (2005)

Hyvärinen, A.: Some extensions of score matching. Comput. Stat. Data Anal. **51**(5), 2499–2512 (2007) https://doi.org/10.1016/j.csda.2006.09.003

Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., Bonneau, R.A.: Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput. Biol. **11**(5), 1004226 (2015) https://doi.org/10.1371/journal.pcbi.1004226

Kassel, L.H., Scealy, J., Bell, B.M.: Score Matching Estimation by Automatic Differentiation (2024). https://cloud.r-project.org/web/packages/scorematchingad/scorematchingad.pdf

Lin, L., Drton, M., Shojaie, A.: Estimation of High-Dimensional graphical models using regularized score matching. Electron. J. Stat. **10**(1), 806–854 (2016) https://doi.org/10.1214/16-EJS1126

Lubbe, S., Filzmoser, P., Templ, M.: Comparison of zero replacement strategies for compositional data with large numbers of zeros. Chemometrics Intellig. Lab. Syst. **210**, 104248 (2021) https://doi.org/10.1016/j.chemolab.2021.104248

Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **15**(12), 550 (2014) https://doi.org/10.1186/s13059-014-0550-8

Lin, H., Peddada, S.D.: Analysis of compositions of microbiomes with bias correction. Nat. Commun. **11**(1), 3514 (2020) https://doi.org/10.1038/s41467-020-17041-7

Maier, M.J.: DirichletReg: Dirichlet regression for compositional data in R. Research Report Series, Vienna University of Economics and Business **125**(125) (2014)

Ma, S., Huttenhower, C., Janson, L.: Compositional differential abundance testing: Defining and finding a new type of health-microbiome associations. bioRxiv, 2024–0604596112 (2024) https://doi.org/10.1101/2024.06.04.596112

Mishra, A.K., Müller, C.L.: Negative binomial factor regression with application to microbiome data analysis. Stat. Med. **41**(15), 2786–2803 (2022) https://doi.org/10.1002/sim.9384

McNulty, R., Sritharan, D., Pahng, S.H., Meisch, J.P., Liu, S., Brennan, M.A., Saxer, G., Hormoz, S., Rosenthal, A.Z.: Probe-based bacterial single-cell RNA sequencing predicts toxin regulation. Nat Microbiol **8**(5), 934–945 (2023) https://doi.org/10.1038/s41564-023-01348-4

Nearing, J.T., Douglas, G.M., Hayes, M.G., MacDonald, J., Desai, D.K., Allward, N., Jones, C.M., Wright, R.J., Dhanani, A.S., Comeau, A.M., *et al.*: Microbiome differential abundance methods produce different results across 38 datasets. Nature communications **13**(1), 342 (2022)

Ostner, J., Carcy, S., Müller, C.L.: tascCODA: Bayesian tree-aggregated analysis of compositional amplicon and single-cell data. Front. Genet. **12**, 766405 (2021) https://doi.org/10.3389/fgene.2021.766405

Obregon-Tito, A.J., Tito, R.Y., Metcalf, J., Sankaranarayanan, K., Clemente, J.C., Ursell, L.K., Zech Xu, Z., Van Treuren, W., Knight, R., Gaffney, P.M., Spicer, P., Lawson, P., Marin-Reyes, L., Trujillo-Villarroel, O., Foster, M., Guija-Poma, E., Troncoso-Corzo, L., Warinner, C., Ozga, A.T., Lewis, C.M.: Subsistence strategies in traditional societies distinguish gut microbiomes. Nat. Commun. **6**(1), 6505 (2015) https://doi.org/10.1038/ncomms7505

Perez, R.K., Gordon, M.G., Subramaniam, M., Kim, M.C., Hartoularos, G.C., Targ, S., Sun, Y., Ogorodnikov, A., Bueno, R., Lu, A., Thompson, M., Rappoport, N., Dahl, A., Lanata, C.M., Matloubian, M., Maliskova, L., Kwek, S.S., Li, T., Slyper, M., Waldman, J., Dionne, D., Rozenblatt-Rosen, O., Fong, L., Dall'Era, M., Balliu, B., Regev, A., Yazdany, J., Criswell, L.A., Zaitlen, N., Ye, C.J.: Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. Science **376**(6589), 1970 (2022) https://doi.org/10.1126/science.abf1970

Quinn, T.P., Erb, I., Richardson, M.F., Crowley, T.M.: Understanding sequencing data as compositions: an outlook and review. Bioinformatics **34**(16), 2870–2878 (2018) https://doi.org/10.1093/bioinformatics/bty175

Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J.C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C.P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T.N., Shalek, A., Shapiro, E., Sharma, P., Shin, J.W., Stegle, O., Stratton, M., Stubbington, M.J.T., Theis, F.J., Uhlen, M., Oudenaarden, A., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., Yosef, N., Human Cell Atlas Meeting Participants: The human cell atlas. Elife **6** (2017) https://doi.org/10.7554/eLife.27041

Scealy, J.L., Hingee, K.L., Kent, J.T., Wood, A.T.A.: Robust score matching for compositional data. Stat. Comput. **34**(2), 93 (2024) https://doi.org/10.1007/s11222-024-10412-w

Silverman, J.D., Roche, K., Holmes, Z.C., David, L.A., Mukherjee, S.: Bayesian multinomial logistic normal models through marginally latent matrix-t processes. Journal of Machine Learning Research **23**(7), 1–42 (2022)

Schaipp, F., Vlasovets, O., Müller, C.: GGLasso - a python package for general graphical lasso computation. J. Open Source Softw. **6**(68), 3865 (2021) https://doi.org/10.21105/joss.03865

Scealy, J.L., Wood, A.T.A.: Score matching for compositional distributions. J. Am. Stat. Assoc. **118**(543), 1811–1823 (2022) https://doi.org/10.1080/01621459.2021.2016422

Te Beest, D.E., Nijhuis, E.H., Möhlmann, T.W.R., Ter Braak, C.J.F.: Log-ratio analysis of microbiome data with many zeroes is library size dependent. Mol. Ecol. Resour. **21**(6), 1866–1874 (2021) https://doi.org/10.1111/1755-0998.13391

Tsilimigras, M.C.B., Fodor, A.A.: Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Ann. Epidemiol. **26**(5), 330–335 (2016) https://doi.org/10.1016/j.annepidem.2016.03.002

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., Gordon, J.I.: The human microbiome project. Nature **449**(7164), 804–810 (2007) https://doi.org/10.1038/nature06244

Tsagris, M., Preston, S., Wood, A.T.A.: Improved classification for compositional data using the α-transformation. J. Classif. **33**(2), 243–261 (2016) https://doi.org/10.1007/s00357-016-9207-5

Wadsworth, W.D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S.A., Vannucci, M.: An integrative bayesian dirichlet-multinomial regression model for the

analysis of taxonomic abundances in microbiome data. BMC Bioinformatics **18**(1), 94 (2017) https://doi.org/10.1186/s12859-017-1516-0

Weistuch, C., Zhu, J., Deasy, J.O., Tannenbaum, A.R.: The maximum entropy principle for compositional data. BMC Bioinformatics **23**(1), 449 (2022) https://doi.org/10.1186/s12859-022-05007-z

Xia, F., Chen, J., Fung, W.K., Li, H.: A logistic normal multinomial regression model for microbiome compositional data analysis. Biometrics **69**(4), 1053–1063 (2013) https://doi.org/10.1111/biom.12079

Yu, S., Drton, M., Shojaie, A.: Generalized score matching for Non-Negative data. J. Mach. Learn. Res. **20** (2019)

Yu, S., Drton, M., Shojaie, A.: Generalized score matching for general domains. Inf inference **11**(2), 739–780 (2022) https://doi.org/10.1093/imaiai/iaaa041

Yu, S., Drton, M., Shojaie, A.: Interaction models and generalized score matching for compositional data. In: Villar, S., Chamberlain, B. (eds.) Proceedings of the Second Learning on Graphs Conference. Proceedings of Machine Learning Research, vol. 231, pp. 1–25 (2024). https://proceedings.mlr.press/v231/yu24a.html

Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., Heath, A.C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J.G., Lozupone, C.A., Lauber, C., Clemente, J.C., Knights, D., Knight, R., Gordon, J.I.: Human gut microbiome viewed across age and geography. Nature **486**(7402), 222–227 (2012) https://doi.org/10.1038/nature11053

Zhou, H., He, K., Chen, J., Zhang, X.: LinDA: linear models for differential abundance analysis of microbiome compositional data. Genome Biol. **23**(1), 95 (2022) https://doi.org/10.1186/s13059-022-02655-5 arXiv:2104.00242 [stat.ME]

Zeng, Y., Pang, D., Zhao, H., Wang, T.: A zero-inflated logistic normal multinomial model for extracting microbial compositions. J. Am. Stat. Assoc., 1–14 (2022) https://doi.org/10.1080/01621459.2022.2044827

Appendix A Connection between a-b power interaction models and PPI models

The class of polynomially tilted pairwise interaction (PPI) models, introduced by Scealy and Wood (2022), is another class of flexible distributions with feature interactions on the simplex. This class includes distributions of the form

$$p_{\boldsymbol{A}^*,\boldsymbol{\nu}}(\boldsymbol{x}) \propto \left(\prod_{i=1}^{p} x_i^{\nu_i}\right) \exp(\boldsymbol{x}^T \boldsymbol{A}^* \boldsymbol{x})$$

with $\boldsymbol{\nu} \succ -1 \in \mathbb{R}^p$, and $\boldsymbol{A}^* \in \mathbb{R}^{p\times p}$ symmetric with $\boldsymbol{A}^* \mathbf{1}_p = 0$. Through $\left(\prod_{i=1}^{p} x_i^{\nu_i}\right) = \exp(\boldsymbol{\nu}^T \log(\boldsymbol{x}))$, it is easy to see that this class of distributions represents a special case of a-b power interaction models (Eq. 2) with $a = 1$ and $b = 0$. Profiling out the last coordinate, i.e. $x_p = 1 - \sum_{i=1}^{p-1} x_i$, leads to an alternative formulation (Scealy et al., 2024), with parameters $\boldsymbol{\nu} \succ -1 \in \mathbb{R}^p$, $\boldsymbol{A}_L \in \mathbb{R}^{(p-1)\times(p-1)}$, and $\boldsymbol{c}_L \in \mathbb{R}^{(p-1)}$:

$$p_{\boldsymbol{A}_L,\boldsymbol{c}_L,\boldsymbol{\nu}}(\boldsymbol{x}) \propto \left(\prod_{i=1}^{p} x_i^{\nu_i}\right) \exp(\boldsymbol{x}^T \boldsymbol{A}_L \boldsymbol{x} + \boldsymbol{c}_L^T \boldsymbol{x}).$$

In particular, the transition between the two forms can be achieved by splitting off the last row and column of $\boldsymbol{A}^* = \begin{pmatrix} \boldsymbol{A}_L^* & \boldsymbol{A}_p^* \\ \boldsymbol{A}_p^{*T} & \boldsymbol{A}_{pp}^* \end{pmatrix}$. Then, $\boldsymbol{A}_{Li,j} = \boldsymbol{A}^*{}_{i,j} - 2\boldsymbol{A}_{pi}^* + \boldsymbol{A}_{pp}^*$ and $\boldsymbol{c}_{Li} = 2(\boldsymbol{A}_{pi}^* - \boldsymbol{A}_{pp}^*)$. Since $\boldsymbol{A}^*$ has one additional parameter, assume $\boldsymbol{A}_{pp}^* = 0$ for the reverse transformation. Then, $\boldsymbol{A}_{pi}^* = \frac{1}{2}\boldsymbol{c}_{Li}$, and $\boldsymbol{A}_{Li,j}^* = \boldsymbol{A}_{Li,j} + \boldsymbol{c}_{Li}$.

Applying the equivalent transformations to an a-b power interaction model with $a = 1$ can help with parameter interpretation, as the matrix $\boldsymbol{A}_L$ usually has full rank.

Appendix B Derivation of the parameters in the quadratic form of the score matching optimizer

This section details the derivation of the parameters $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ in the quadratic formulation of the score matching loss (Eq. 8) and explains their block structure shown in Eq. 11. The elements of $\boldsymbol{g}$ can be directly derived from the second derivative of $\log p(\boldsymbol{x})$ (Eq. 10):

$$\begin{aligned}
\boldsymbol{g}_{\mathbf{K},j} \equiv \frac{1}{n}\sum_{i=1}^{n} &\left[\partial_j \tilde{h}_j\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{a-1} + (a-1)\tilde{h}_j\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{a-2}\right] {\boldsymbol{X}^{(i)}}^a \\
&+ a\tilde{h}_j\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{2a-2} \boldsymbol{e}_{j,p} - a\tilde{h}_j\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{a-1} {X_p^{(i)}}^{a-1} \boldsymbol{e}_{p,p}, \\
\boldsymbol{g}_{\mathbf{K},p} \equiv \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{p-1} &\left[-\partial_k \tilde{h}_k\left(\boldsymbol{X}^{(i)}\right) {X_p^{(i)}}^{a-1} + (a-1)\tilde{h}_k\left(\boldsymbol{X}^{(i)}\right) {X_p^{(i)}}^{a-2}\right] {\boldsymbol{X}^{(i)}}^a
\end{aligned}$$

$$+ a\tilde{h}_k\left(\boldsymbol{X}^{(i)}\right) {X_p^{(i)}}^{2a-2}\boldsymbol{e}_{p,p} - a\tilde{h}_k\left(\boldsymbol{X}^{(i)}\right) {X_k^{(i)}}^{a-1}{X_p^{(i)}}^{a-1}\boldsymbol{e}_{k,p},$$

$$\boldsymbol{g}_{\boldsymbol{\eta_0},j} \equiv \frac{1}{n}\sum_{i=1}^{n} -\partial_j\tilde{h}_j\left(\boldsymbol{X}^{(i)}\right){X_j^{(i)}}^{b-1} - (b-1)\tilde{h}_j\left(\boldsymbol{X}^{(i)}\right){X_j^{(i)}}^{b-2} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}^{(i)}_{\boldsymbol{\eta_1},j},$$

$$\boldsymbol{g}_{\boldsymbol{\eta_0},p} \equiv \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{p-1} \partial_k\tilde{h}_k\left(\boldsymbol{X}^{(i)}\right){X_p^{(i)}}^{b-1} - (b-1)\tilde{h}_k\left(\boldsymbol{X}^{(i)}\right){X_p^{(i)}}^{b-2} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}^{(i)}_{\boldsymbol{\eta_1},p},$$

$$\boldsymbol{g}_{\boldsymbol{\eta_1},j} \equiv \frac{1}{n}\sum_{i=1}^{n} -\partial_j\tilde{h}_j\left(\boldsymbol{X}^{(i)}\right)y{X_j^{(i)}}^{b-1} - (b-1)\tilde{h}_j\left(\boldsymbol{X}^{(i)}\right)y{X_j^{(i)}}^{b-2} = \frac{1}{n}\sum_{i=1}^{n}y\boldsymbol{g}^{(i)}_{\boldsymbol{\eta_1},j},$$

$$\boldsymbol{g}_{\boldsymbol{\eta_1},p} \equiv \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{p-1} \partial_k\tilde{h}_k\left(\boldsymbol{X}^{(i)}\right)y{X_p^{(i)}}^{b-1} - (b-1)\tilde{h}_k\left(\boldsymbol{X}^{(i)}\right)y{X_p^{(i)}}^{b-2} = \frac{1}{n}\sum_{i=1}^{n}y\boldsymbol{g}^{(i)}_{\boldsymbol{\eta_1},p}.$$

Further, the elements of $\boldsymbol{\Gamma}$ follow from the first derivative of $\log p(\boldsymbol{x})$ (Eq. 9) and have the same structure as in Yu et al. (2024):

$$\boldsymbol{\Gamma}_{\mathbf{K}} \equiv \begin{bmatrix} \boldsymbol{\Gamma}_{\mathbf{K},1} & \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{\Gamma}_{\mathbf{K},(1,p)} \\ \mathbf{0} & \boldsymbol{\Gamma}_{\mathbf{K},2} & \cdots & \mathbf{0} & \boldsymbol{\Gamma}_{\mathbf{K},(2,p)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Gamma}_{\mathbf{K},p-1} & \boldsymbol{\Gamma}_{\mathbf{K},(p-1,p)} \\ \boldsymbol{\Gamma}^{\top}_{\mathbf{K},(1,p)} & \boldsymbol{\Gamma}^{\top}_{\mathbf{K},(2,p)} & \cdots & \boldsymbol{\Gamma}^{\top}_{\mathbf{K},(p-1,p)} & \boldsymbol{\Gamma}_{\mathbf{K},p} \end{bmatrix} \in \mathbb{R}^{p^2\times p^2},$$

with each block of size $p \times p$, and

$$\boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta_i}} \equiv \begin{bmatrix} \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_i},1} & \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_i},(1,p)} \\ \mathbf{0} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_i},2} & \cdots & \mathbf{0} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_i},(2,p)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_i},p-1} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_i},(p-1,p)} \\ \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_i},(p,1)} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_i},(p,2)} & \cdots & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_i},(p,p-1)} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_i},p} \end{bmatrix} \in \mathbb{R}^{p^2\times p} \; for \; i \in \{1,2\},$$

with each block a vector of size p, and

$$\boldsymbol{\Gamma}_{\boldsymbol{\eta_i}} \equiv \begin{bmatrix} \gamma_{\boldsymbol{\eta_i},1} & 0 & \cdots & 0 & \gamma_{\boldsymbol{\eta_i},(1,p)} \\ 0 & \gamma_{\boldsymbol{\eta_i},2} & \cdots & 0 & \gamma_{\boldsymbol{\eta_i},(2,p)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \gamma_{\boldsymbol{\eta_i},p-1} & \gamma_{\boldsymbol{\eta_i},(p-1,p)} \\ \gamma_{\boldsymbol{\eta_i},(1,p)} & \gamma_{\boldsymbol{\eta_i},(2,p)} & \cdots & \gamma_{\boldsymbol{\eta_i},(p-1,p)} & \gamma_{\boldsymbol{\eta_i},p} \end{bmatrix} \in \mathbb{R}^{p\times p} \; for \; i \in \{1,2\},$$

, and

$$\boldsymbol{\Gamma}_{\boldsymbol{\eta_0},\boldsymbol{\eta_1}} \equiv \begin{bmatrix} \gamma_{\boldsymbol{\eta_0},\boldsymbol{\eta_1},1} & 0 & \cdots & 0 & \gamma_{\boldsymbol{\eta_0},\boldsymbol{\eta_1},(1,p)} \\ 0 & \gamma_{\boldsymbol{\eta_0},\boldsymbol{\eta_1},2} & \cdots & 0 & \gamma_{\boldsymbol{\eta_0},\boldsymbol{\eta_1},(2,p)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \gamma_{\boldsymbol{\eta_0},\boldsymbol{\eta_1},p-1} & \gamma_{\boldsymbol{\eta_0},\boldsymbol{\eta_1},(p-1,p)} \\ \gamma_{\boldsymbol{\eta_0},\boldsymbol{\eta_1},(1,p)} & \gamma_{\boldsymbol{\eta_0},\boldsymbol{\eta_1},(2,p)} & \cdots & \gamma_{\boldsymbol{\eta_0},\boldsymbol{\eta_1},(p-1,p)} & \gamma_{\boldsymbol{\eta_0},\boldsymbol{\eta_1},p} \end{bmatrix} \in \mathbb{R}^{p\times p}.$$

These blocks have the following specific forms. For $j = 1, \ldots, p-1$,

$$\begin{aligned} \boldsymbol{\Gamma}_j &\equiv \begin{bmatrix} \boldsymbol{\Gamma}_{\mathbf{K},j} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_0},j} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_1},j} \\ \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_0},j}^\top & \gamma_{\boldsymbol{\eta_0},j} & \gamma_{\boldsymbol{\eta_0},\boldsymbol{\eta_1},j} \\ \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_1},j}^\top & \gamma_{\boldsymbol{\eta_0},\boldsymbol{\eta_1},j}^\top & \gamma_{\boldsymbol{\eta_1},j} \end{bmatrix} \\ &\equiv \frac{1}{n}\sum_{i=1}^{n} \tilde{h}_j\left(\boldsymbol{X}^{(i)}\right) \begin{bmatrix} {X_j^{(i)}}^{a-1}\boldsymbol{X}^{(i)a} \\ -{X_j^{(i)}}^{b-1} \\ -y{X_j^{(i)}}^{b-1} \end{bmatrix} \begin{bmatrix} {X_j^{(i)}}^{a-1}\boldsymbol{X}^{(i)a} \\ -{X_j^{(i)}}^{b-1} \\ -y{X_j^{(i)}}^{b-1} \end{bmatrix}^\top, \\ \boldsymbol{\Gamma}_p &\equiv \begin{bmatrix} \boldsymbol{\Gamma}_{\mathbf{K},p} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_0},p} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_1},p} \\ \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_0},p}^\top & \gamma_{\boldsymbol{\eta_0},p} & \gamma_{\boldsymbol{\eta_0},\boldsymbol{\eta_1},p} \\ \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_1},p}^\top & \gamma_{\boldsymbol{\eta_0},\boldsymbol{\eta_1},p}^\top & \gamma_{\boldsymbol{\eta_1},p} \end{bmatrix} \\ &\equiv \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{p-1} \tilde{h}_k\left(\boldsymbol{X}^{(i)}\right) \begin{bmatrix} {X_p^{(i)}}^{a-1}\boldsymbol{X}^{(i)a} \\ -{X_p^{(i)}}^{b-1} \\ -y{X_p^{(i)}}^{b-1} \end{bmatrix} \begin{bmatrix} {X_p^{(i)}}^{a-1}\boldsymbol{X}^{(i)a} \\ -{X_p^{(i)}}^{b-1} \\ -y{X_p^{(i)}}^{b-1} \end{bmatrix}^\top, \\ \boldsymbol{\Gamma}_{(j,p)} &\equiv \begin{bmatrix} \boldsymbol{\Gamma}_{\mathbf{K},(j,p)} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_0},(j,p)} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_1},(j,p)} \\ \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_0},(p,j)} & \gamma_{\boldsymbol{\eta_0},(j,p)} & \gamma_{\boldsymbol{\eta_0},\boldsymbol{\eta_1},(j,p)} \\ \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta_1},(p,j)} & \boldsymbol{\gamma}_{\boldsymbol{\eta_0},\boldsymbol{\eta_1},(p,j)} & \gamma_{\boldsymbol{\eta_1},(j,p)} \end{bmatrix} \\ &\equiv -\frac{1}{n}\sum_{i=1}^{n} \tilde{h}_j\left(\boldsymbol{X}^{(i)}\right) \begin{bmatrix} {X_j^{(i)}}^{a-1}\boldsymbol{X}^{(i)a} \\ -{X_j^{(i)}}^{b-1} \\ -y{X_j^{(i)}}^{b-1} \end{bmatrix} \begin{bmatrix} {X_p^{(i)}}^{a-1}\boldsymbol{X}^{(i)a} \\ -{X_p^{(i)}}^{b-1} \\ -y{X_p^{(i)}}^{b-1} \end{bmatrix}^\top. \end{aligned}$$

Appendix C Scaling score matching elements to approximate Box-Cox transformations

As described in Section 2.4.2, the power transformation used for a-b power interaction models (Eqs. 4 and 2) bears striking resemblance to the Box-Cox transformation $\frac{1}{\phi}(x^\phi - 1)$. Both transformations are not equivalent though due to the subtraction of 1 in the Box-Cox transformation. This difference causes the a-b power interaction transformation to lose one key property of the Box-Cox transformation - its asymptotic approximation of the logarithm as ϕ approaches 0.

Looking at the density of the covariate-extended a-b power interaction model makes this disparity clear:

$$p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^{a} + \frac{1}{b}(\boldsymbol{\eta}_0 + \boldsymbol{y}\boldsymbol{\eta}_1)^{\top}\boldsymbol{x}^{b}\right)$$

For the terms concerning $\boldsymbol{\eta_0}$ and $\boldsymbol{\eta_1}$, the subtraction of 1 in the Box-Cox transformation is not dependent on x and can therefore be absorbed into the normalizing constant. For the interaction term, replacing $\boldsymbol{x}^a$ with the Box-Cox transformation in Eqs. 2 or 4 would introduce a scaling factor of order $1/a^2$ instead of $1/a$, leading to a discontinuity of the estimated $\boldsymbol{K}$ when approaching the log-log model, for which the convention $\frac{1}{2a} \equiv 1$ is used (Yu et al., 2024).

We counteract this effect by introducing scaling factors of $1/a$ and $1/a^2$ on the components of $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ (Eq. 11), based on the matrix multiplication $\boldsymbol{\theta}^{\top}\boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta}$ from Eq. 8. In particular, we scale $\boldsymbol{\Gamma}_{\boldsymbol{K}}$ by a factor of $\frac{1}{a^2}$ and $\boldsymbol{\Gamma}_{\boldsymbol{K},\boldsymbol{\eta_0}}, \boldsymbol{\Gamma}_{\boldsymbol{K},\boldsymbol{\eta_1}}$, and $\boldsymbol{g}_{\boldsymbol{K}}$ by a factor of $\frac{1}{a}$ each. This leads to a smooth transition in the estimation of $\boldsymbol{K}$ when $\phi \to 0$, and also holds for general a-b power interaction models without covariates (Yu et al., 2024).

We showcase the effectiveness of our scaling approach with an example on the scRNA-seq data of SLE patients and healthy controls Perez et al. (2022). For simplicity, we estimate the whole dataset through the covariate-less a-b power interaction model without differentiating between the two groups, use no regularization on the off-diagonal entries of $\boldsymbol{K}$, and always replace zeros with a value of 0.5. Without the scaling factor, the pattern of the estimated $\boldsymbol{K}$ approaches the log-solution ($\phi = 0$), but the scale of the entries is not the same (Figure E14, left column). On the other hand, the entries of $\boldsymbol{\eta}$ approach the log-solution also in magnitude (Figure E14, right column). For increasing values of ϕ, both the pattern and magnitude of $\boldsymbol{K}$ and $\boldsymbol{\eta}$ gradually diverge, as the power transformation gradually distorts the composition differently.

The median entry of the ratio $\boldsymbol{K}_{\boldsymbol{\phi=0}}/\boldsymbol{K}_{\boldsymbol{\phi=\phi'}}$ also does not approach 1 as $\phi' \to 0$ (Figure E15, bottom right). Looking at the components of $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$, one can see that the median entry of the above ratio follows a log-linear trend for larger values of ϕ, but not for smaller exponents if the component is associated with $\boldsymbol{K}$ (Figure E15, other panels). The scaling factors introduced above correct this trend, such that the ratio is log-linear across the full spectrum of ϕ. This causes the estimated $\boldsymbol{K}$ to approach the solution for $\phi = 0$ in magnitude (Figure E15, bottom right) without impacting the estimated interaction pattern (Figure E14, middle column) or the estimation of $\boldsymbol{\eta}$ (Figure E14, right column).

When combining regularization and power transforms, the dependency between ϕ and the scale of entries in $\boldsymbol{K}$ will lead to different optimal regularization strengths for different exponents (Figure E16a). In fact, a larger exponent and therefore larger scale of $\boldsymbol{K}$ will require smaller values of λ_1 to cover the whole range between $\boldsymbol{K}$ with full support and a diagonal $\boldsymbol{K}$ (Figure E16b). Therefore, the range of values for λ_1 should always be adapted to the current data and power transform.

Appendix D Testing for differential abundance without feature interactions

We also compared the methods on simulated data without feature interactions to show the suitability of `cosmoDA` if no significant feature associations are present. To this end, we applied the a-b power model solution with $a = b = 0$ and $\lambda_1 = 2$ to the dataset from Perez et al. (2022), resulting in ground truth parameters of $K_B = \mathbf{0}^{11\times 11}$, and $\boldsymbol{\eta}_{0,B}$ as shown in Figure E3. We used the same setup as before to select differentially abundant cell types and effect sizes and again chose $n = 100$ and $n = 1000$, simulating five replicates for each of the 30 scenarios as described above.

If no significant feature interactions were simulated, `cosmoDA` and CompDA showed similar overall performance as before, while the MCC of ANCOM-BC and Dirichlet regression significantly improved (Figures D1a, E8). This improvement was due to a reduction in falsely discovered effects by these methods (Figure D1a), which shows that the high FDR of ANCOM-BC and Dirichlet regression in the previous simulation were caused by secondary effects due to feature interactions. In terms of power, all methods showed similar strength as before (Figures D1c, E10). Nevertheless, `cosmoDA` was the only model to consistently produce a FDR close to the nominal level, while CompDA was not able to avoid false discoveries if the effect was placed on the abundant cell type T4 (Figure E9). The superior performance of `cosmoDA` in this case was due to the fact the the data was simulated by an a-b power interaction model, which is not used by the other methods.

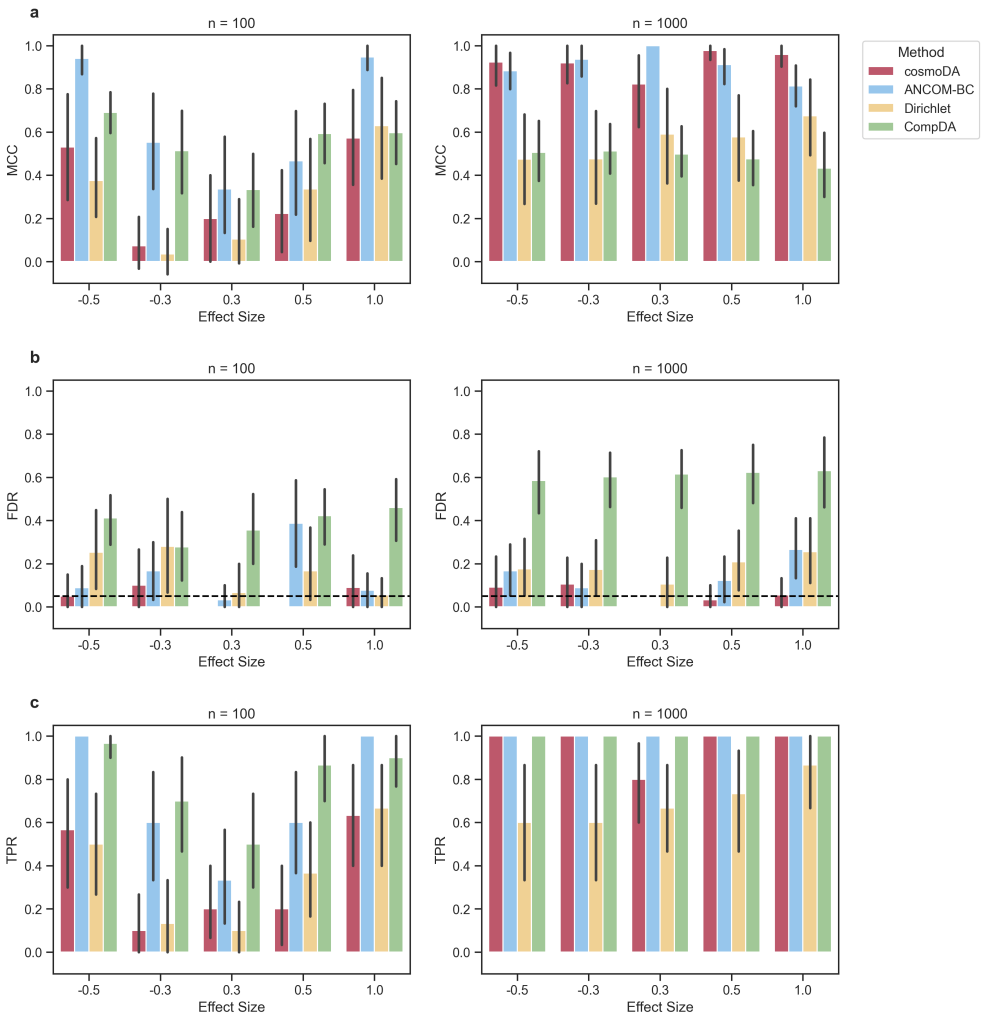


Fig. D1 Performance comparison for recovering differentially abundant features across different scenarios, $K = 0$**.** (a) Matthews' correlation coefficient. (b) False discovery rate. The dashed line shows the nominal FDR for all methods. (c) True positive rate (power).

Appendix E Supplementary Figures

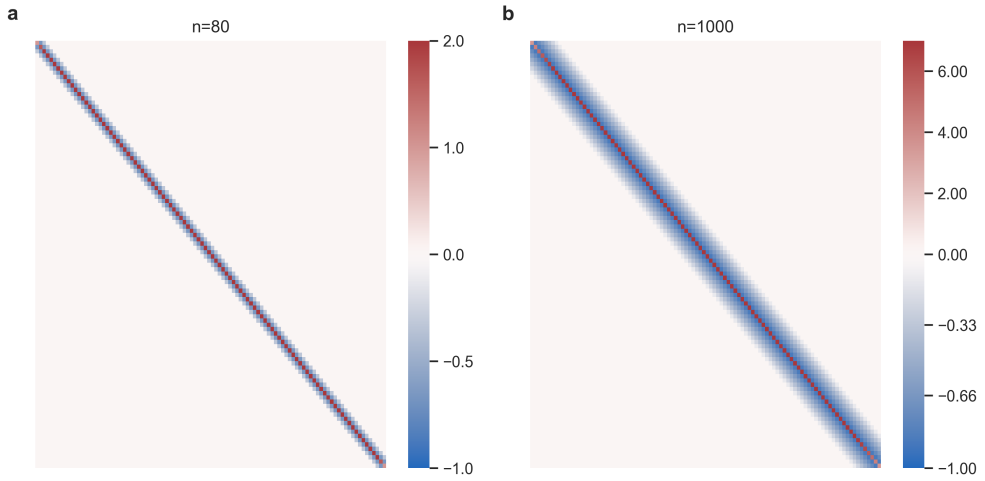


Fig. E2 Interaction matrices used for data generation in the benchmark testing recovery of $\boldsymbol{K}$ (Section 3.1). (a) $n = 80$, (b) $n = 1000$.

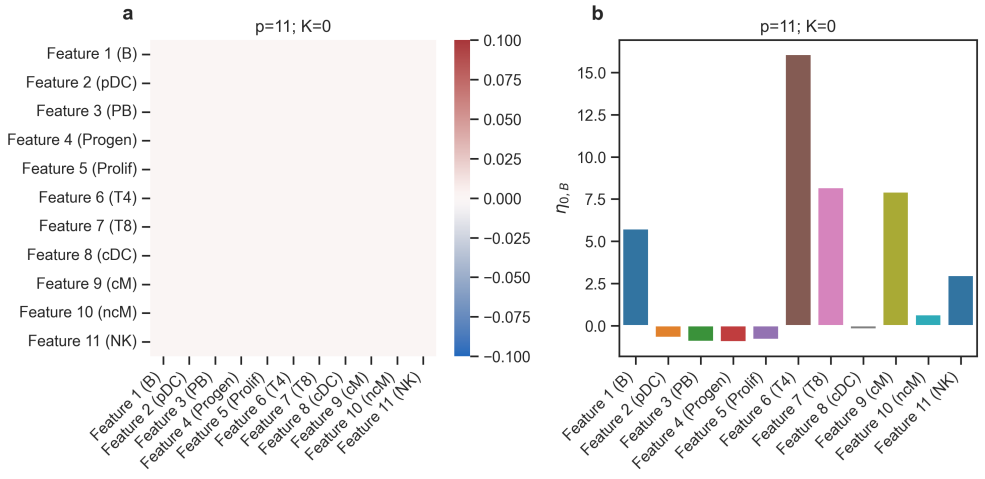


Fig. E3 Data generation parameters used for the differential abundance testing benchmark (Section 3.2), $K = 0$**.** (a) Interaction matrix ($\boldsymbol{K}_B$). (b) Location vector ($\boldsymbol{\eta}_{0,B}$).

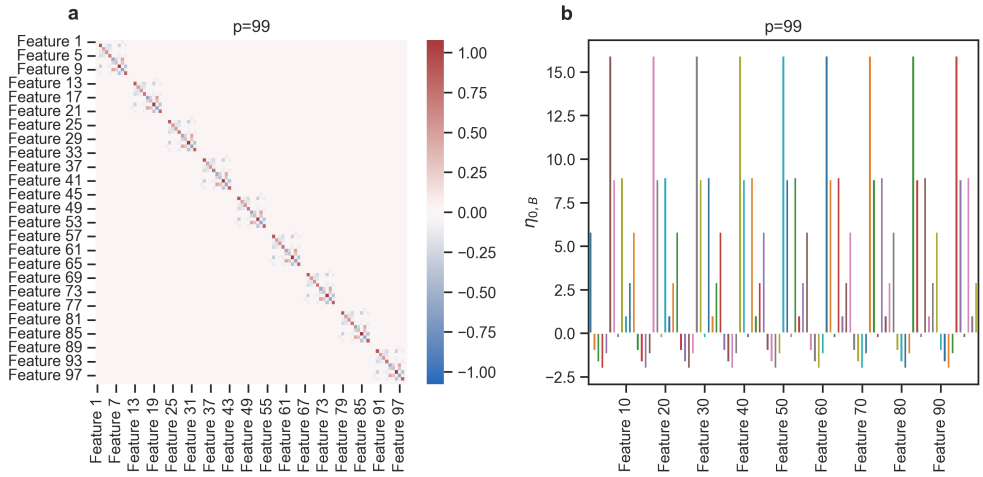


Fig. E4 Data generation parameters used for the differential abundance testing benchmark (Section 3.2), $p = 99$. (a) Interaction matrix ($\boldsymbol{K}_B$). (b) Location vector ($\boldsymbol{\eta}_{0,B}$).

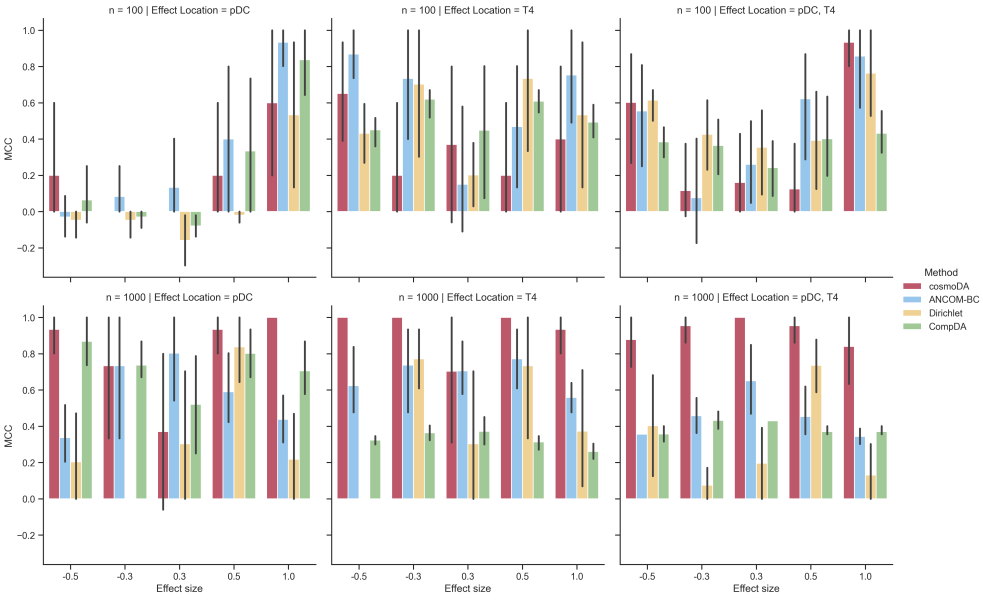


Fig. E5 Detailed breakdown of Matthews' correlation coefficient for the differential abundance testing benchmark (Section 3.2), $p = 11$.

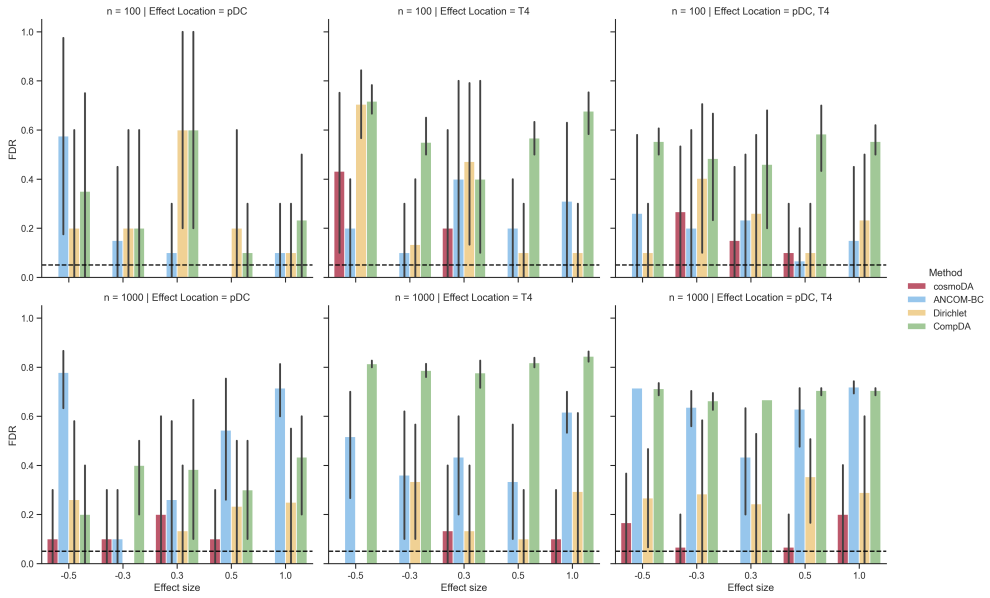


Fig. E6 Detailed breakdown of false discovery rate for the differential abundance testing benchmark (Section 3.2), $p = 11$. The dashed lines denote the nominal FDR for all methods.

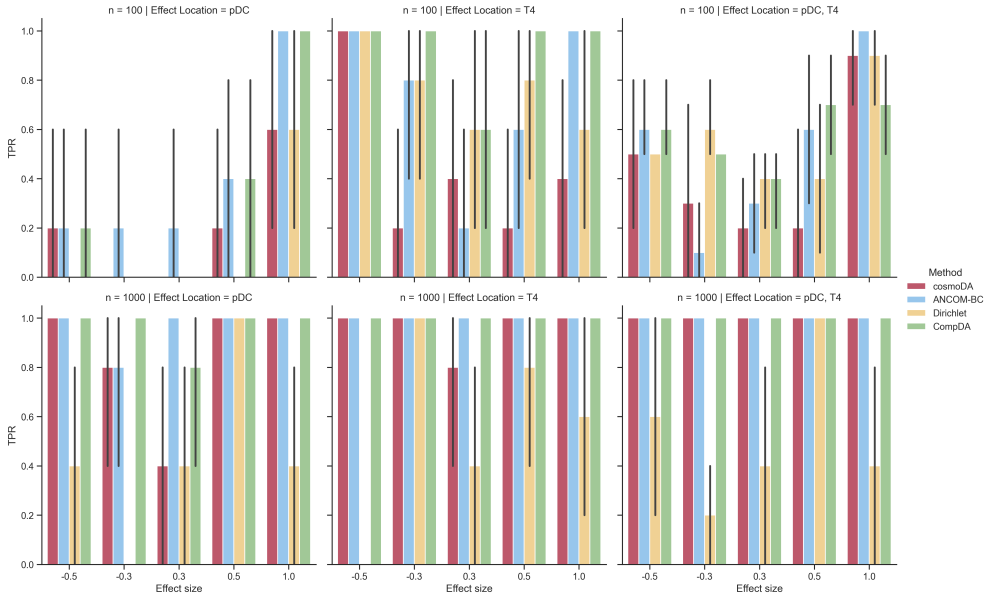


Fig. E7 Detailed breakdown of power (true positive rate) for the differential abundance testing benchmark (Section 3.2), $p = 11$.

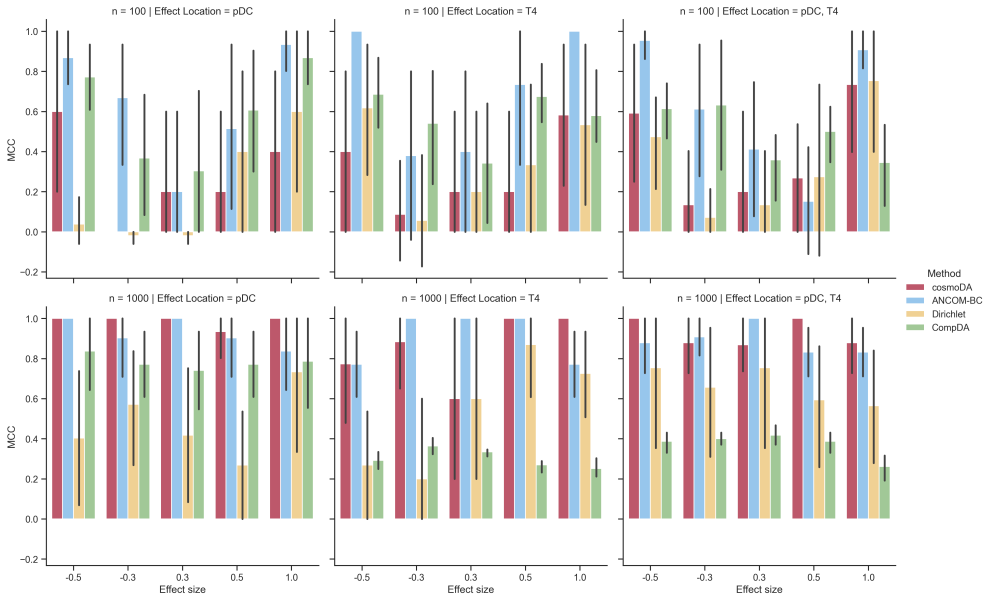


Fig. E8 Detailed breakdown of Matthews' correlation coefficient for the differential abundance testing benchmark (Section 3.2), $K = 0$.

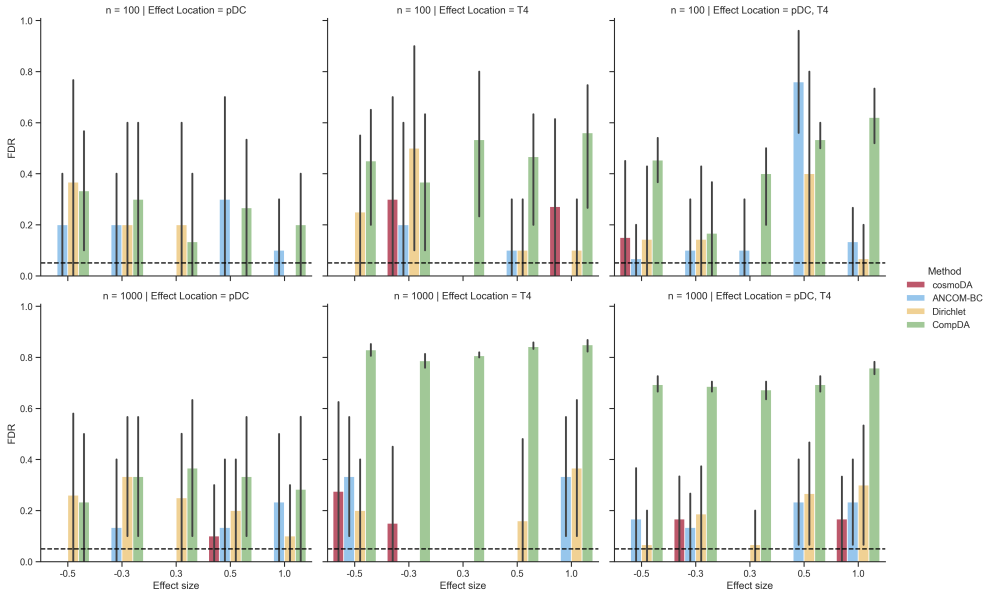


Fig. E9 Detailed breakdown of false discovery rate for the differential abundance testing benchmark (Section 3.2), $K = 0$. The dashed lines denote the nominal FDR for all methods.

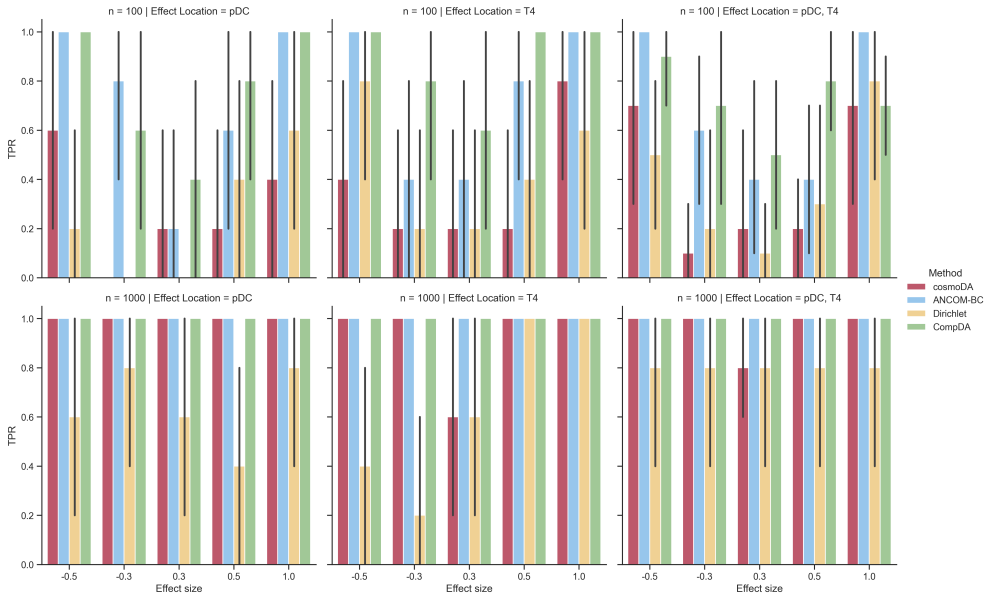


Fig. E10 Detailed breakdown of power (true positive rate) for the differential abundance testing benchmark (Section 3.2), $K = 0$.

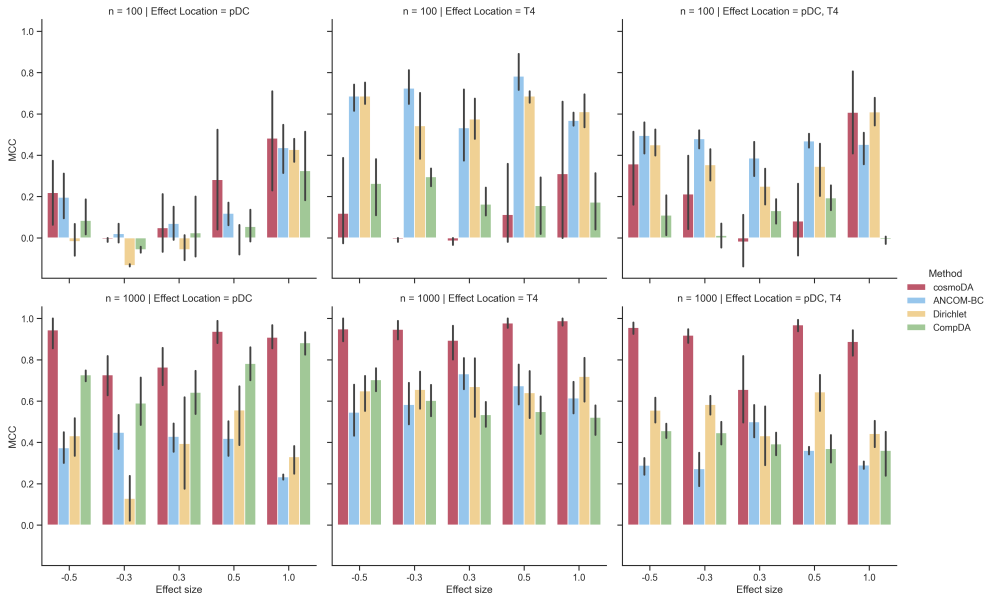


Fig. E11 Detailed breakdown of Matthews' correlation coefficient for the differential abundance testing benchmark (Section 3.2), $p = 99$.

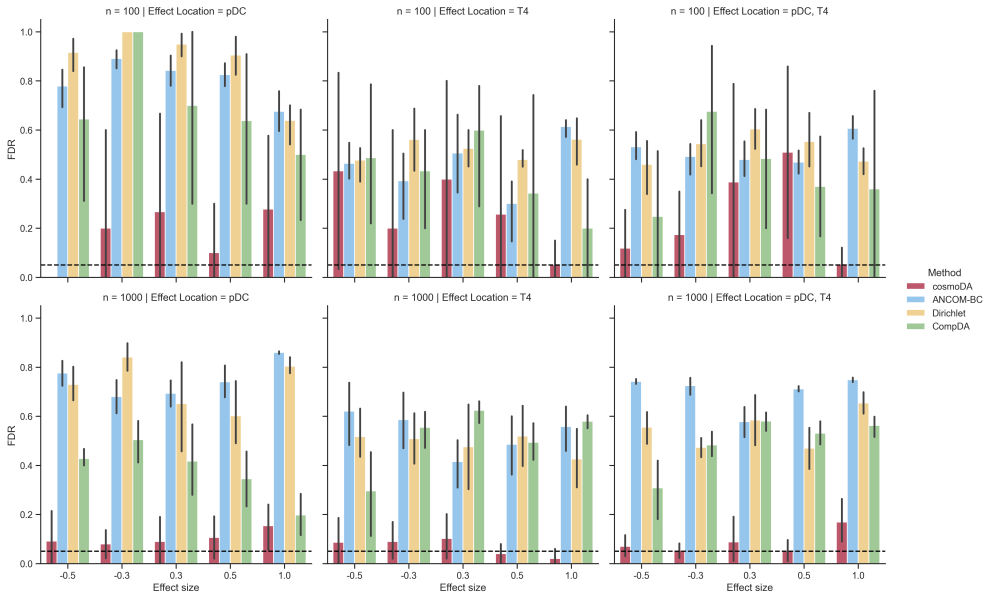


Fig. E12 Detailed breakdown of false discovery rate for the differential abundance testing benchmark (Section 3.2), $p = 99$. The dashed lines denote the nominal FDR for all methods.

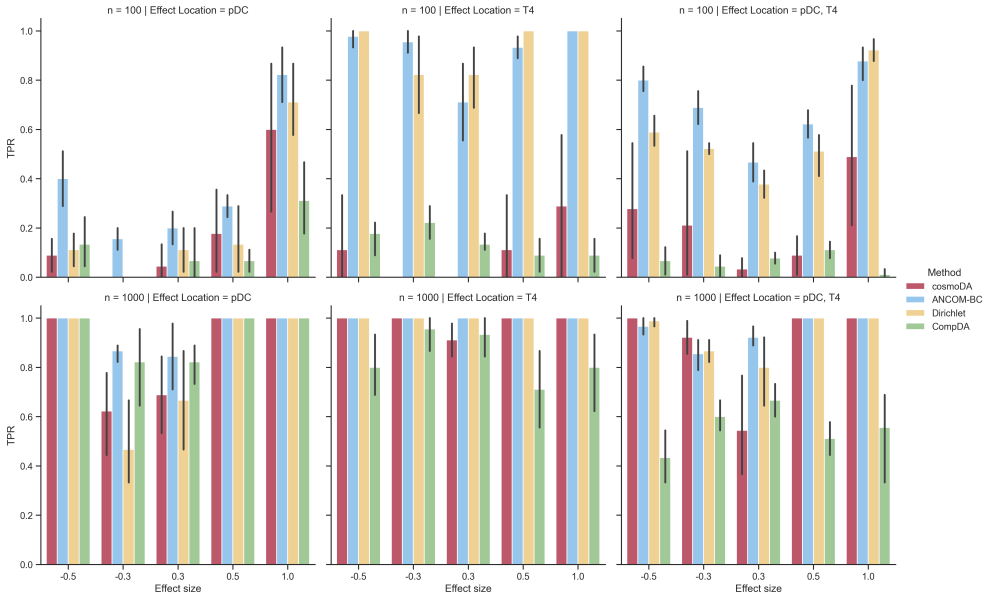


Fig. E13 Detailed breakdown of power (true positive rate) for the differential abundance testing benchmark (Section 3.2), $p = 99$.

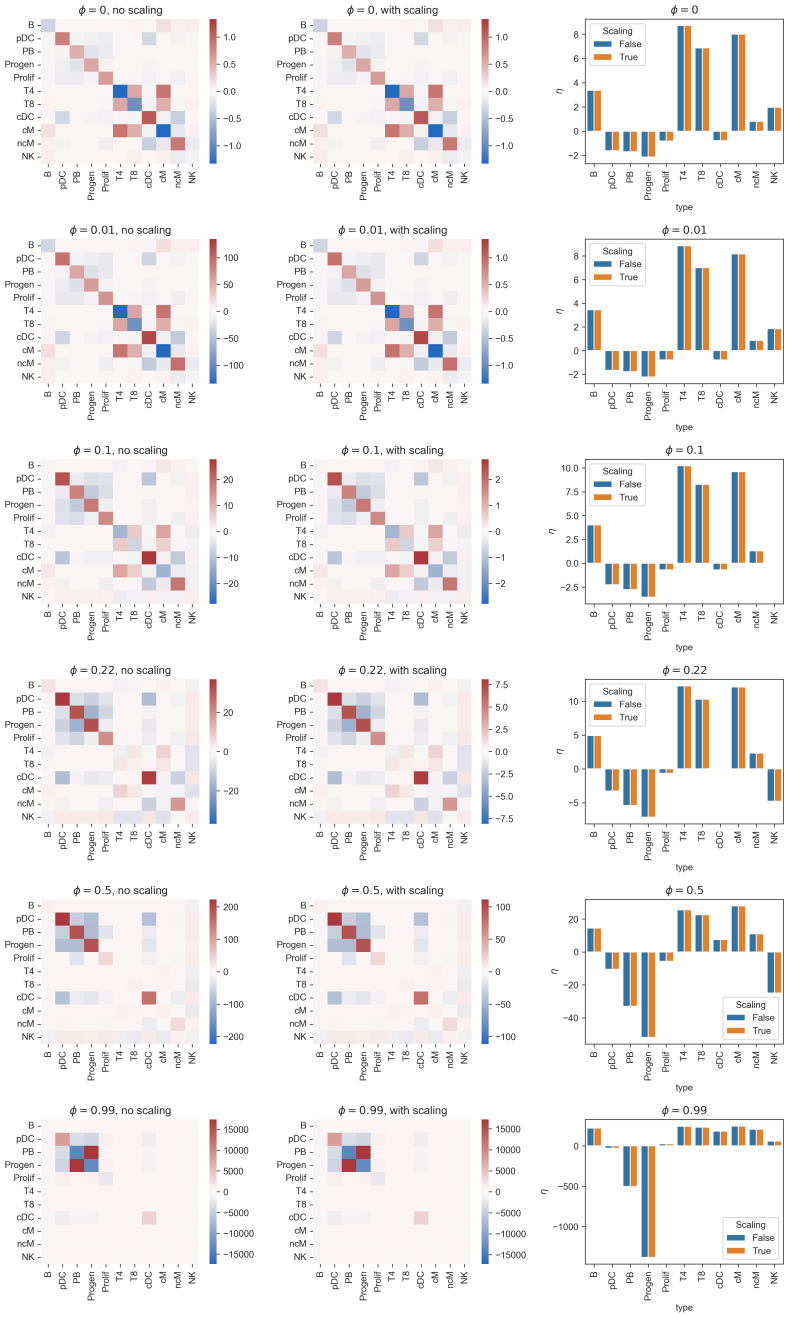


Fig. E14 Impact of scaling Γ and g on the estimation of K and η. Results shown for the SLA scRNA-seq data Perez et al. (2022). Rows show selected values of the exponent ϕ in the power transformation. Left column: Values of $\boldsymbol{K}$ without scaling. Middle column: Values of $\boldsymbol{K}$ with scaling. Right column: Values of $\boldsymbol{\eta}$ with and without scaling.

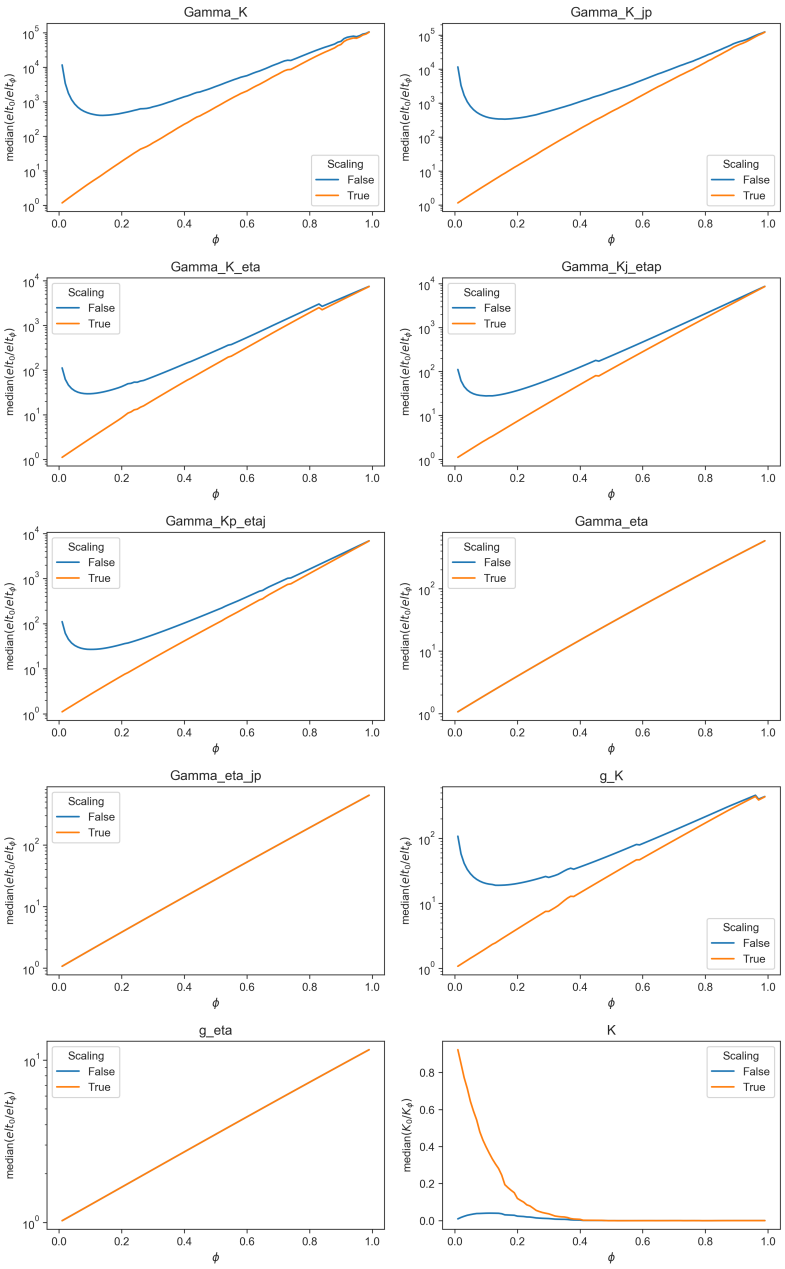


Fig. E15 Impact of scaling factor for power transforms on the score matching parameters Γ and g and the interaction matrix K. All plots except bottom right show the median entry of $E_{\phi=0}/E_{\phi=\phi'}$ for E being one of the score matching elements in Eq. 11. Bottom right: Same quantity for the estimated interaction matrix $\boldsymbol{K}$.

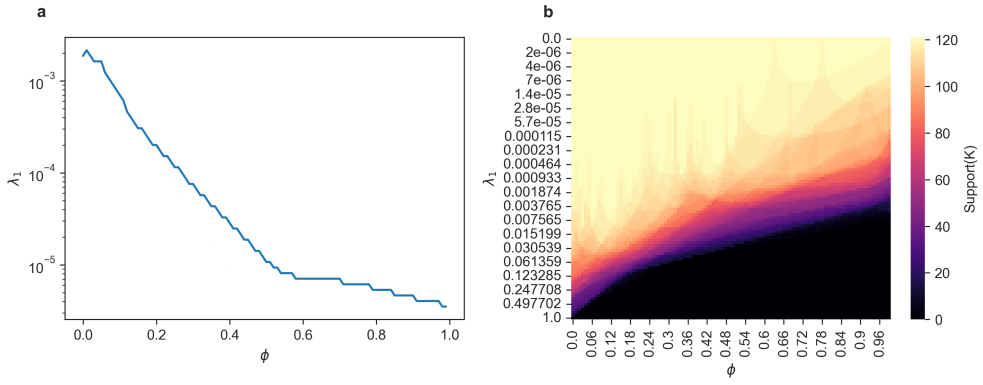


Fig. E16 Relationship between power and regularization strength for the SLA scRNA-seq data Perez et al. (2022). (a) Value of λ_1 selected through cross validation in relation to exponent ϕ of the power transform. (b) Number of nonzero entries in $\boldsymbol{K}$ for every λ_1 and ϕ.

C. Best practices for analysis of HTS data

C.1. BacSC: A general workflow for bacterial single-cell RNA sequencing data analysis

Contributing article

Ostner, J., Kirk, T., Olayo-Alarcon, R., Thöming, J., Rosenthal, A. Z., Häussler, S., et al. (2024). BacSC: A general workflow for bacterial single-cell RNA sequencing data analysis. *bioRxiv, 2024-06.* doi: `https://doi.org/10.1101/2024.06.22.600071`

Replication code

Source code for this contribution has been deposited on Github at `https://github.com/bio-datascience/BacSC`. Supplemental data can be downloaded from zenodo (`https://zenodo.org/records/12189002`).

Copyright information

Author contributions

J.O. and C.L.M. designed the structure and individual steps of the BacSC pipeline, and conceived improvements to existing methods. T.K., J.G.T. and S.H. generated data containing P. aeruginosa and E. coli with ProBac-seq, A.Z.R. provided all datasets from B. subtilis. J.O. implemented the pipeline and conducted all applications and tests. J.O., T.K., R.O.A., J.G.T., and A.Z.R. analyzed the results from BacSC in a biological context, R.O.A. further performed analysis of the Co-PATHOgenex data. J.O. wrote the manuscript with help from all other authors. All authors read and approved the manuscript.

BacSC: A general workflow for bacterial single-cell RNA sequencing data analysis

Johannes Ostner[1,2,8*], Tim Kirk[3], Roberto Olayo-Alarcon[2], Janne Gesine Thöming[3,4], Adam Z. Rosenthal[5], Susanne Häussler[3,4,6], Christian L. Müller[1,2,7*]

[1]Computational Health Center, Helmholtz Munich, Ingolstädter Landstraße 1, 85764, Neuherberg, Germany.
[2]Institut für Statistik, Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, 80539, Munich, Germany.
[3]Institute for Molecular Bacteriology, TWINCORE, Centre for Experimental and Clinical Infection Research, Hannover, Germany.
[4]Department of Clinical Microbiology, Rigshospitalet, Copenhagen, Denmark.
[5]Department of Microbiology and Immunology, University of North Carolina, Chapel Hill, NC, USA.
[6]Department of Molecular Bacteriology, Helmholtz Centre for Infection Research, Braunschweig, Germany.
[7]Center for Computational Mathematics, Flatiron Institute, 162 Fifth Avenue, New York, 10010, NY, USA.
[8]Lead contact.

*Corresponding author(s). E-mail(s):
johannes.ostner@stat.uni-muenchen.de;
christian.mueller@helmholtz-munich.de;

Abstract

Bacterial single-cell RNA sequencing has the potential to elucidate within-population heterogeneity of prokaryotes, as well as their interaction with host systems. Despite conceptual similarities, the statistical properties of bacterial single-cell datasets are highly dependent on the protocol, making proper processing essential to tap their full potential. We present BacSC, a fully data-driven

computational pipeline that processes bacterial single-cell data without requiring manual intervention. BacSC performs data-adaptive quality control and variance stabilization, selects suitable parameters for dimension reduction, neighborhood embedding, and clustering, and provides false discovery rate control in differential gene expression testing. We validated BacSC on a broad selection of bacterial single-cell datasets spanning multiple protocols and species. Here, BacSC detected subpopulations in *Klebsiella pneumoniae*, found matching structures of *Pseudomonas aeruginosa* under regular and low-iron conditions, and better represented subpopulation dynamics of *Bacillus subtilis*. BacSC thus simplifies statistical processing of bacterial single-cell data and reduces the danger of incorrect processing.

Keywords: bacterial single-cell RNA sequencing, phenotypic heterogeneity, statistical analysis, data processing, computational pipeline, data thinning, synthetic data generation, scanpy

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized genetic analysis of eukaryotic cell compendia by allowing researchers to extract individual cells' gene expression profiles and obtain new insights on intracellular mechanisms, as well as the structure and dynamics within entire populations of cells [1–3]. These advances have led, among others, to a better understanding of immune responses [4], disease progression [5], or advancements in drug development [6]. Consequently, similar insights into microbial heterogeneity are expected from scRNA-seq of bacterial populations, opening up new avenues for assessing antimicrobial resistance, evolutionary pathways, or within-population differences in response to external conditions [7]. In addition, bacterial scRNA-seq yields new ways to analyze interactions between the isogenic microbiome and host systems, for example in toxin regulation [8, 9], formation of metabolic niches [10], and the analysis of microbial spatial heterogeneity [11].

Applying scRNA-seq technologies to bacteria has however proven to be challenging, e.g. due to low overall transcript abundance, the short half-life of bacterial mRNA, and difficulties in cell lysis due to sturdier cell walls [12–15]. Recently, multiple protocols have been developed that enable scRNA-seq of bacteria on larger scales by tackling these challenges in different ways [13, 14, 16–19]. For example, ProBac-seq [18] uses a library of oligonucleotide probes to target mRNAs, while BacDrop [13] uses a two-stage cell barcoding procedure to increase cell numbers.

Datasets from scRNA-seq contain gene expression counts for each UMI (unique molecular identifier) and are typically sparse, high-dimensional and noisy, requiring specialized methods and particular care in their statistical processing to obtain biologically meaningful representations [20, 21]. This process has been extensively discussed for eukaryotic cells, leading to well-documented benchmarks [22, 23], best practices [24–26], and methods to select adequate hyperparameters [27–29] for each step of the statistical analysis pipeline. For bacterial scRNA-seq, no such guidelines exist yet, prompting the use of default parameters and methods without prior assessment of their statistical validity and suitability for the data at hand. This may, however, lead to suboptimal or even flawed representations of the data, which can severely impact the quality of biological insights gained from downstream analyses.

Each step in a typical statistical processing pipeline developed for the analysis of eukaryotic scRNA-seq [24, 26] poses new statistical challenges when applied to bacterial scRNA-seq data:

- In quality control, differences in sparsity and sequencing depth have to be accounted for when filtering out low-quality genes or cells [30].
- Variance stabilization is a crucial step to ensure comparability for all sequenced cells, but scaling the data to a common sequencing depth and the choice of an imputation value for zero replacement must be done with the statistical properties of the data in mind [21, 22].
- The number of principal components used for low-dimensional data representation, as well as the number of neighbors and minimal distance used in UMAP embeddings, are hyperparameters that are commonly chosen in a heuristic fashion, but have a significant impact on downstream analysis and visual representation of the data [27, 31].
- The resolution parameter in cell type clustering is also often determined by visual trial-and-error procedures [32].
- Finally, recent studies show that differential expression testing between cell types suffers from a double-dipping issue that inflates the false discovery rate [29] if not accounted for.

In this study, we address these challenges by developing a standard workflow for processing bacterial scRNA-seq gene expression data that does not require the selection of modeling choices or manual tuning of parameters. We introduce BacSC, a computational pipeline for automatic processing of scRNA-seq data that is applicable to datasets generated by various bacterial scRNA-seq protocols. BacSC reevaluates the validity of methods used in each of the steps outlined above in the context of bacterial scRNA-seq, adjusts methods if necessary, and automatically chooses suitable hyperparameters in a data-driven way. To this end, BacSC provides tools for data integration and quality control of bacterial scRNA-seq data, and performs a simple, yet powerful variance stabilizing transform that is suitable for scRNA-seq data with varying sequencing depth and high zero inflation. Using techniques from data thinning [31, 33] and knockoff generation [27, 34], BacSC is able to select suitable parameters and perform dimensionality reduction, neighborhood embedding, and cell-type clustering without requiring user intervention. BacSC also offers FDR control for differential expression testing of bacterial scRNA-seq data through contrasting p-values with synthetic null data[29, 35].

To validate the steps taken in BacSC, we compared the statistical properties of 13 datasets generated with ProBac-seq [18, 36] and BacDrop [13], emphasizing their low sequencing depth, high zero inflation,

and differences in marginal gene distribution. As a proof of concept, BacSC was able to distinguish the same cell types as previously shown through analysis with default or manually chosen parameters for all datasets with known biological structure. BacSC additionally showed improved ability to describe the transitional nature of cell competence in *B. subtilis*, was able to give a more clear distinction of cells expressing mobile genetic elements in *K. pneumoniae*, and discovered new cellular subpopulations in *K. pneumoniae* and *P. aeruginosa*. When applied to a combination of *P. aeruginosa* cells grown under regular and iron-reduced conditions, BacSC was able to simultaneously integrate cells from both conditions based on their gene expression profiles and detect differential expression of genes related to iron acquisition.

BacSC is available as a modular framework in Python that seamlessly integrates into the scanpy [37] workflow and allows for direct downstream analysis with other tools from the scverse [38]. BacSC is available on GitHub (https://github.com/bio-datascience/BacSC).

2 Results

2.1 Explorative comparison of bacterial scRNA-seq technologies reveals differences in key statistical properties

To ensure the cross-platform and cross-species applicability of BacSC, we gathered a total of 13 bacterial scRNA-seq datasets that were generated with two different sequencing protocols, ProBac-seq [18, 36], and BacDrop [13] (see section 3). The datasets encompass five bacterial species (*Pseudomonas aeruginosa, Bacillus subtilis, Klebsiella pneumoniae, Escherichia coli, Enterococcus faecium*), further distinguished by strain, growth environment, or treatment condition (Table 1).

Dataset	Species/strain	Condition	Protocol	Source
Pseudomonas_balanced_PB	*P. aeruginosa* PAO1	balanced growth	ProBac-seq	This study
Pseudomonas_li_PB	*P. aeruginosa* PAO1	Low Iron environment	ProBac-seq	This study
Ecoli_balanced_PB	*E. coli* MG1655	balanced growth	ProBac-seq	This study
Bsub_minmed_PB	*B. subtilis* 168	minimal media	ProBac-seq	McNulty et al. [18]
Bsub_damage_PB	*B. subtilis* 168	DNA damage induced by Mitomycin C	ProBac-seq	This study
Bsub_MPA_PB	*B. subtilis* 168	MPA energy stress	ProBac-seq	This study
Klebs_antibiotics_BD	*K. pneumoniae* MGH66	6 samples, treated with one of 3 antibiotics (2 samples each): Meropenem, Gentamicin, Ciprofloxacin	BacDrop	Ma et al. [13]
Klebs_untreated_BD	*K. pneumoniae* MGH66	Untreated culture (2 samples)	BacDrop	Ma et al. [13]
Klebs_BIDMC35_BD	*K. pneumoniae* BIDMC35	Untreated culture	BacDrop	Ma et al. [13]
Klebs_4species_BD	*K. pneumoniae* MGH66	Untreated culture	BacDrop	Ma et al. [13]
Ecoli_4species_BD	*E. coli* 10ß	Untreated culture	BacDrop	Ma et al. [13]
Efaecium_4species_BD	*E. faecium* EnGen0052	Untreated culture	BacDrop	Ma et al. [13]
Pseudomonas_4species_BD	*P. aeruginosa* PAO1	Untreated culture	BacDrop	Ma et al. [13]

Table 1 Description of datasets used to benchmark BacSC. All datasets are named by the convention *species_condition_protocol*. Datasets from ProBac-seq are marked with the suffix "_PB", datasets from BacDrop are marked with "_BD"

The number of genes per dataset was mostly dependent on the species (Figure 2A), and ranged between 5,572 (*P. aeruginosa*) and 2,350 (*E. faecium*). The sequencing depth per cell was highly dependent on the sequencing method, with data from BacDrop showing a median sequencing depth between 2 and 43, while all datasets generated with ProBac-seq had at least a median sequencing depth of 150 (Figure 2B). In contrast, datasets generated with BacDrop generally encompassed a higher number of cells (median 9,936) than datasets from ProBac-seq (median 3,773).

After filtering out cells with abnormally low or high expression and genes without reads in more than one cell (See section 2.2), both protocols could be easily distinguished by the number of genes detected, with all datasets from ProBac-seq encompassing at least 2,922 genes, while datasets from BacDrop contained a maximum of 2,500 genes (Figure 2C, Table E1). This was in part due to the subsetting to 2,500 highly variable genes, which was only performed on the *Klebs_antibiotics_BD, Klebs_untreated_BD*, and *Klebs_BIDMC35_BD* datasets. The BacDrop data from the four species comparison comprised a much lower numbers of genes (628 - 1606) without selection of highly variable genes. The number of cells generally differed more within the BacDrop data (103 - 48,511), while the ProBac-seq datasets had much more stable cell numbers (1,910 - 13,801; Figure 2C, Table E1).

BacDrop only detected between 24 and 47 unique genes per cell on average, while ProBac-seq covered at least 49 genes for each cell in every dataset (Figure 2D). Consequently, ProBac-seq had less zero entries in the filtered read count matrices, with zeroes making up between 86% and 97% of all entries, while BacDrop showed zero inflation numbers between 95% and 99.2% (Figure 2E). After quality control, we observed similar discrepancies between protocols in sequencing depth. ProBac-seq not only covered more genes per cell, but was also able to capture more transcripts, with median sequencing depths ranging from 103 to 794.5. BacDrop datasets only had a median sequencing depth of 45 or less after quality control (Figure 2F; Table E1). We therefore reasoned that the usage of multiple probes per gene and subsequent

aggregation through max-pooling in ProBac-seq (see Methods, [36]) leads to higher genome coverage and sequencing depth for each cell.

2.2 Description of the BacSC pipeline

At its core, statistical processing of scRNA-seq data extracts information from raw transcriptome reads by filtering, normalization, dimension reduction, and clustering steps [24, 26]. BacSC selects suitable methods and automates the choice of hyperparameters for each step without the need for manual intervention (except for quality control; Figure 1). Section 2.2 briefly describes each step, while we give more detailed descriptions in the "QUANTIFICATION AND STATISTICAL ANALYSIS" part of the STAR methods.

First, the data is subjected to quality control to filter out barcodes with abnormally low or high gene expression (Figure 1A). Because our exploratory analysis showed that bacterial single-cell data differs heavily in terms of average sequencing depth, number of expressed genes, and zero inflation, this step is highly dependent on the experimental protocol used. Therefore, BacSC leaves this step as the only point where manual intervention is necessary, but provides tools for outlier detection through median absolute deviation (MAD) statistics [30] and aggregating probe-based data from ProBac-seq. As with eukaryotic scRNA-seq data, the main data object after quality control in each dataset is a $n \times p$-dimensional count table X, containing the read counts of p features for n cells.

Next, the read count data must be normalized and scaled. Because bacterial scRNA-seq data shows greatly reduced sequencing depth and increased zero inflation compared to eukaryotic scRNA-seq, special care has to be taken in this step [39, 40]. BacSC first scales each cell individually to have the same number of reads, and subsequently log-transforms the data. The pseudocount introduced in this step is gene-specific [22], with overdispersion parameters calculated through sctransform [41] (Figure 1A). Finally, each gene is scaled to have zero mean and unit variance over all cells.

After variance stabilization, the data is reduced to a lower-dimensional representation by singular value decomposition (SVD) on the data. The embedding dimensionality k in this step of the scRNA-seq processing workflow is often set manually, e.g. by finding an "elbow" in the plot of SVD loadings [25]. BacSC instead uses a count-splitting approach to find a good value for k, which was described by Neufeld et al. [31]. For this, the raw counts after quality control are split into a training and test dataset, and the variance-stabilizing transform is applied to both datasets. Then, the latent dimensionality k with minimal reconstruction error between the k-dimensional embedding of the training data and the full test data is chosen (Figures 1A, B1).

UMAP (Uniform Manifold Approximation and Projection) plots [42] are a popular tool for two-dimensional visualization of scRNA-seq data to preserve the local structure and point out global differences in higher-dimensional data. The algorithm is largely dependent on three parameters - the latent dimensionality k, the number of neighbors $n_{neighbors}$ considered for each cell, as well as the minimal distance min_{dist} between points. These parameters are often adjusted manually until a satisfactory picture arises. To eliminate this manual step, BacSC uses the negative-control approach described by scDEED [27] to determine the latter two latent parameters. scDEED calculates a reliability score - the correlation between the distance vectors from a cell to its neighbors before and after UMAP embedding - and compares them to the distribution of contrast scores on a randomly permuted dataset (Figures 1A, B1). It then selects the parameter combination for which the amount of cells with abnormally low reliability scores is minimized.

Cell clusters in scRNA-seq data are typically detected through the the Louvain [43] and Leiden [44] algorithms. Both algorithms aim to maximize the modularity of partition over all cells with respect to a resolution parameter res. Once again, this parameter is usually chosen manually to fit the structure observed in the UMAP or PCA embeddings. Computational determination of a feasible resolution parameter that robustly detects cell clusters without creating too many subclusters is, however, not straightforward. BacSC uses the train and test dataset obtained from count splitting and introduces a new gap statistic based on the difference in modularity between two clusterings on the test data - one calculated on the train data and one assigned randomly. Maximizing this gap statistic allows to find a value for res for which the obtained clustering on the train data also generalizes well to the structure of the test data Figures 1A, B2).

Bacterial single-cell sequencing allows to characterize heterogeneity within bacterial populations in unprecedented detail. The discovery of subpopulations and the description and interpretation of different cell types in bacterial populations is therefore still at an early stage. To characterize previously unknown cell types, automatic selection of signature genes for each cluster is often achieved through differential expression (DE) testing [24]. For this task, BacSC provides capabilities for DE testing that takes the

recently popularized problem of "double dipping" for DE testing of cell types into account [29, 45, 46]. In short, using the same information (gene expression) to define a clustering as well as the subsequently determining DE genes to characterize these clusters results in an inflated false discovery rate (FDR). BacSC solves this issue by adapting the ClusterDE method [29] for FDR control. Due to the highly sparse nature of bacterial single-cell data, BacSC uses a modified version of scDesign2 [47] to generate the synthetic null data. Further, BacSC also adapts ClusterDE to achieve better results for highly uneven cluster proportions (Figures 1B, , B3).

To validate our pipeline, we applied BacSC to all datasets described in Table 1. For quality control, we manually set dataset-specific filtering parameters on minimal sequencing depth and MAD cutoff (Table E2), based on visual inspection of the distribution of sequencing depth and number of unique genes per cell. After variance stabilization, we further reduced the *Klebs_antibiotics_BD, Klebs_untreated_BD*, and *Klebs_BIDMC35_BD* datasets to 2,500 highly variable genes based on their standardized variances [48]. All other steps of BacSC do not require any manual intervention, and were thus performed automatically. The determined data distribution, as well as parameters for latent dimensionality, number of neighbors, minimal distance, and clustering resolution are shown in Table E2.

2.3 BacSC uncovers new biological structures in datasets obtained from different bacterial scRNA-seq protocols

2.3.1 Transitions between cellular states in *B.subtilis* are pronounced by BacSC

To show the validity of the transformations and parameters selected in BacSC, we first investigated the *Bsub_minmed_PB* dataset (Figures 3, D18). This data was generated by [18] to validate the ProBac-seq method. The original analysis with default parameters in Seurat [48] discovered four distinct subpopulations with multiple subclusters and different functionality. In the first two dimensions of the PCA embedding suggested by BacSC, three larger subpopulations were immediately apparent (Figure 3A), while a fourth cluster with only 20 cells emerged in the UMAP embedding with BacSC's selected parameters (Figure 3B). Clustering with the automatically determined resolution resulted in five cell type clusters (Figure 3B).

Because of the "double-dipping" issue described above, DE testing produced large numbers for genes with very small p-values for each cell type (Figure D18I). Counteracting this through the p-value correction in BacSC revealed characteristic genes for each cell type (Figure 3E-G), but only the two smallest clusters (3 and 4) had genes significant at a FDR level of $\alpha = 0.05$ (Figure D18J, Table E4).

Cell type 4 showed increased expression of many sporulation genes (*spoIVA, spoVID, spoIID*), while the marker genes in cell type 3 contained many genes associated with cell competence (*comFA, comGD, comGB, comGA, comGC, comFC*). These subpopulations were also found as clusters 9, respectively 6/8 in [18]. Cell type 0 contained cells with very low sequencing depths (Figure 3C, D), and many genes were significantly underexpressed at an FDR level of 0.1 (Table E4). The genes with the highest contrast scores for this cell type partially overlapped with genes found in clusters 0 and 3 in the original publication. Similarly, cell type 1 contained many upregulated genes at an FDR of 0.2. For cell type 2, many structural flagella components (*fliY, fliD, fliK, fliI, fliT*) were among the genes with the highest contrast scores, but only differentially expressed at an FDR level of 0.26. The region containing cell types 1 and 2 from BacSC therefore corresponds to clusters 1, 2, 3, and 5 from [18].

Notably, the UMAP from BacSC showed continuous streams of cells between the cell types, especially between cell types 0, 1, and 3 (Figure 3B), which were not visible in the original analysis [18]. We suspected these cells to be in a transitional phase between two cell states. The development of competent cells (cell type 3) is known to be procedural [49], which explains the transition of cells in and out of this cell type.

2.3.2 BacSC shows clear differences in response of *K. pneumoniae* to different antibiotics

To showcase the applicability of BacSC to data from different bacterial scRNA-seq protocols, we revisited an analysis of six samples of *Klebsiella pneumoniae* generated with BacDrop [13]. The *Klebs_antibiotics_BD* dataset contains two replicates for each of three antibiotic treatments, ciprofloxacin, meropenem, and gentamicin.

Despite the high sparsity of the data (99.2%, Table E1), BacSC was able to successfully integrate all six samples. The first two principal components already showed heterogeneity in the data in the form of three clear subpopulations (Figure 4A). This was enhanced through the UMAP plot and data clustering (Figure 4B), which revealed two major clusters of cells that split up into two, respectively three cell

types, and three small cell clusters. For all cell types, a subset of genes was differentially expressed at FDR levels of 0.2 or lower (Table E5).

The cell types contained in the largest cellular subpopulation (0, 1, and 2) almost perfectly matched the separation by antibiotics shown in Figure 4D. Within these clusters, cells from both samples were distributed evenly, suggesting no residual batch effects. Cell types 3 and 5 made up all cells in the second large subpopulation, which contained a higher number of unique expressed genes than the rest of the dataset (Figure 4C). Both of these clusters showed significant differential expression of IS903B transposase-related genes (*RS09075, RS22855*), which matches the subpopulation of mobile genetic elements (MGE) described by [13]. Contrary to the original analysis, this subpopulation separated more from the bulk of the cells in BacSC's UMAP embedding (Figure 4B). The small subpopulations (Cell types 4, 6, 7) were all characterized by a few genes that were barely expressed in other cells.

2.4 Processing with BacSC discovers a distinct response of *P. aeruginosa* to a low-iron environment

2.4.1 Bacterial cell types of exponentially grown *P. aeruginosa* are similar in growth conditions with differing iron availability

We next tested if BacSC could recover environment-specific microbial cell types from bacterial cultures grown under different external conditions. For this, we investigated the *Pseudomonas_balanced_PB* and *Pseudomonas_li_PB* datasets. Both datasets contain cells from *P. aeruginosa* in exponential growth in minimal media, and sequenced with ProBac-seq. For the first sample, cells were grown in regular minimal media (MOPS with 10 µM $FeSO_4$), while for the second sample, bacteria were exposed to a mild iron limitation (0.5 µM $FeSO_4$), which resembles a growth condition mimicking competition between host and pathogen for the essential trace element during infection.

We first processed each dataset individually with BacSC. The diagnostic plots for both datasets (D22, D23) showed that normalized sequencing depths, as well as latent dimensionality, neighborhood embedding, and clustering resolution parameters found by BacSC were very similar. The PCA and UMAP embeddings for both datasets also showed similar patterns (Figures 5A, B, C6A, B, C7A, B). The sequencing depth vs. genome coverage plots (Figures 5D, E) revealed that in both populations, a subset of cells had lower coverage at high sequencing depths. This subgroup was identified as cluster 1 in the cell type clustering. Both datasets further contained two larger subpopulations (cell types 0 and 2), and one smaller cluster (cell type 3).

The lower-coverage cell types in both datasets were characterized by 51 and 82 genes respectively, that were differentially expressed at an FDR of 0.05 (Tables E6, E7) when compared to the rest of the population. Of the 95 genes differentially expressed in either of the two datasets, 38 genes appeared in both, including 22 genes encoding components of the 30S and 50S subunits of the ribosome (*rpsA, rpsB, rplQ, rpsKD, rplFO, rplDWBCP, rpmC, rplEN, rpsJ, rpsG, rplJ, rplK, rpsRI*, Figures C6E-G, C7E-G), indicating increased translation activity. Cell type 3 also showed considerable overlap between DE genes at the 5%-level. Here, all 22 genes that were DE in the balanced growth sample were also among the 34 genes detected in the low-iron culture. Many of these genes encode the R-type pyocin R2 (*PA0617, PA0618, PA0619, PA0620, PA0622, PA0623, PA0640*, Figures C6E-G, C7E-G), a phage tail-like bacteriocin that specifically targets and kills competing bacteria by puncturing their cell membranes [50, 51]. For cell type 2, which contained cells with a large number of expressed genes, a large number of genes was detected to be DE at an FDR of 0.05, with underexpressed ribosomal genes showing the highest contrast scores, complementary to the set of DE genes in the low-coverage cell type. The remaining cell type 0 contained cells with low sequencing depth and showed no statistically significant DE genes.

2.4.2 Combined data processing allows for the detection of genes related to iron acquisition

To analyze the differences between the cell populations from balanced and low-iron growth conditions, we created a combined dataset by concatenating the raw count matrices of both experiments. Processing with BacSC revealed a similar common structure as in the individual datasets (Figures Figure 5C, F, C8A-D), confirming the similarities detected in the previous section. While the R2 pyocin cluster (cell type 5) showed good mixing between both conditions, the cell populations with high expression of ribosomal genes distinctly separated and were even clustered into different cell types (2 and 3, (Figure 5C)). Additionally, a new cell type (cluster 4) emerged in the combined dataset, which was not detected in either of the individual datasets. Similar to cell type 0, this cluster showed reduced expression of ribosomal genes (*rplF,*

rplP, rplD, rplB), as well as genes encoding for ATP-synthase and the TCA cycle component succinate dehydrogenase (*atpA, atpD, atpH, sdhA, sdhC*, Figure C8E-G), suggesting a low energy state. For cell types 0 and 1, a within-cluster shift of cells by condition was also visible (Figure 5J). As in the individual data set analyses, marker genes for all cell types except cell type 1 were detected by BacSC at FDR levels smaller than 0.2.

Plotting the cell type proportions for each sample showed that cell types 2 and 3 almost exclusively contained cells from one condition, while the other cell types showed no notable changes in proportionality between the balanced and low-iron conditions (Figure 5M). We confirmed this visual result by differential abundance testing with scCODA [52] and detected cell types 2 and 3 as differentially abundant at an FDR level of 0.2.

Finally, we examined the differences in gene expression between cells from both growth conditions. For this, we first performed DE testing between the balanced growth and low-iron cell populations with a Wilcoxon rank-sum test. Since this test setup does not suffer from double-dipping, we used the Benjamini-Hochberg correction [53] to account for multiple comparisons, revealing 186 genes with corrected p-values of less than 0.05. To verify our findings, we used bulk sequencing results from the Co-PATHOgenex study [54], also testing differential expression between cells grown in balanced and iron-reduced conditions. Of note, in this study an abrupt iron limitation was artificially induced by the addition of the iron chelator 2,2'-bipyridine shortly before harvest. We compared the gene set found by BacSC on the bacterial scRNA-seq data with three gene sets detected on the Co-PATHOgenex data with different DE tests - the method described in the Co-PATHOgenex paper, a logistic regression model, and DESeq2 [55], each at a significance level of 0.05. The gene set from BacSC had good overlap with the gene sets found in bulk data, as 42 of the 186 genes were detected by at least one other DE test, and the intersection of all four gene sets contained 20 genes (Figure 5K). Furthermore, 26 of the 42 genes detected in the bulk data were among the top 50 genes with the lowest adjusted p-values in the DE test on the bacterial scRNA-seq data (Table E3). Investigating the gene expression levels and function of these 42 genes, we found most of them to be overexpressed in the low-iron sample (Figure 5G-I, L). Furthermore, most of these genes (e.g. *PA4514, icmP, phuR*) are known to be related to iron reception (Table E3).

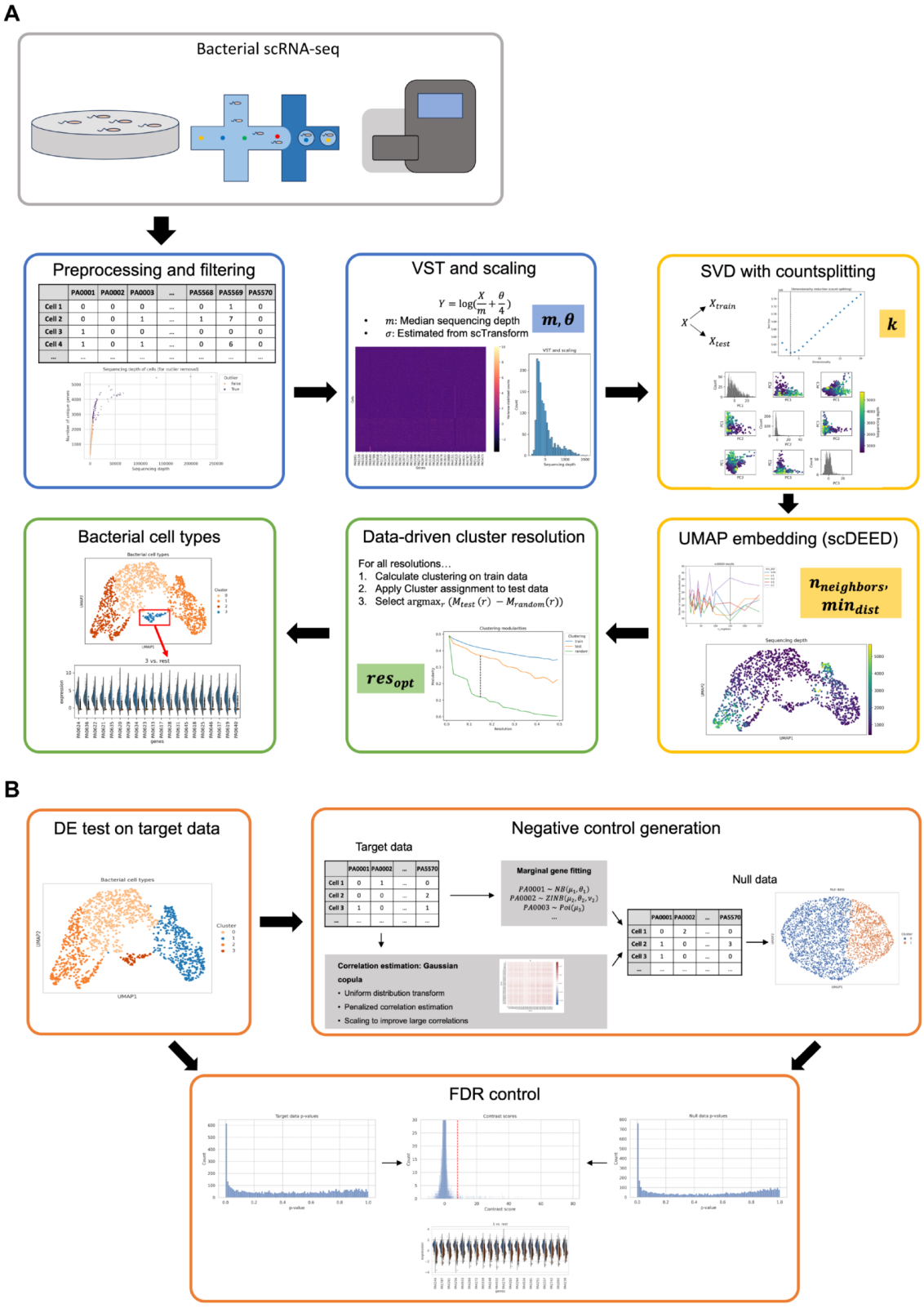


Fig. 1 Conceptual visualization of the BacSC pipeline. (A) Bacterial single-cell RNA sequencing produces a table of read counts. Sterting with this table, BacSC first filters out outlier cells, before performing a variance stabilizing transform (blue boxes). Next, the latent data dimensionality is determined through count splitting, and suitable parameters for UMAP visualizations are determined by scDEED (yellow boxes). Finally, BacSC determines an adequate resolution for clustering, and is able to discover bacterial cell types (green boxes). The colored rectangles show the most important key parameters calculated in the respective step of BacSC. **(B)** For differential expression testing, BacSC generates synthetic null data with the same marginal distributions as the target data through a Gaussian copula approach. P-values of a DE test on the target data are then contrasted with p-values on the synthetic null data to obtain differentially expressed genes at a desired FDR level.

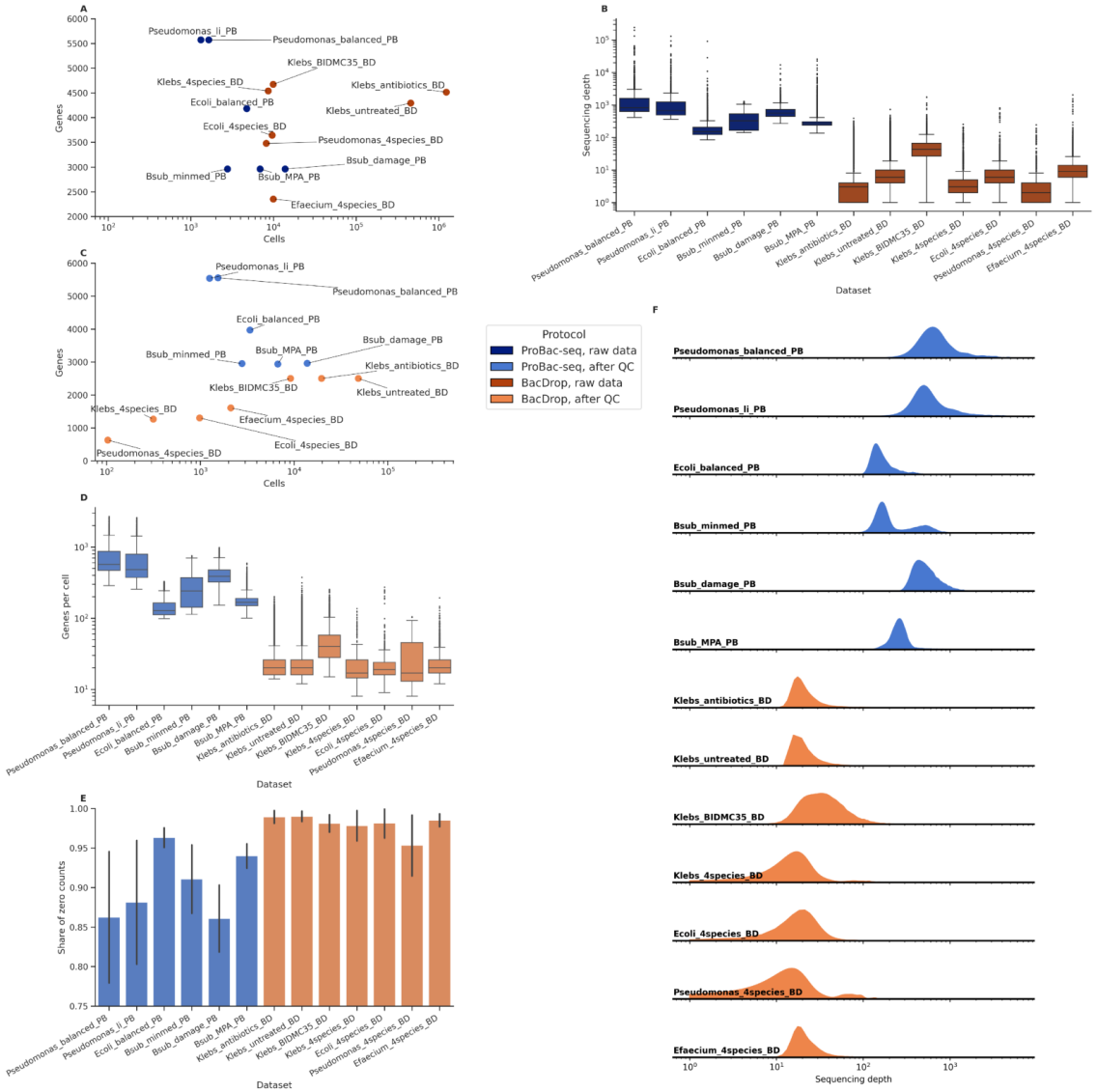


Fig. 2 Explorative comparison of bacterial scRNA-seq technologies reveals differences in key statistical properties. **(A)** Number of genes and cells before quality control. **(B)** Sequencing depth per cell before quality control. The box depicts the 25% and 75% quartiles of the data, as well as the median; whiskers extend to 1.5 times the interquartile range of the data. **(C)** Number of genes and cells after quality control. **(D)** Number of expressed genes per cell after quality control. The box depicts the 25% and 75% quartiles of the data, as well as the median; whiskers extend to 1.5 times the interquartile range of the data. **(E)** Share of zero counts over all cells in the raw data matrices after quality control. Errorbars show the empirical standard deviation. **(F)** Density plots of sequencing depth for each dataset after quality control.

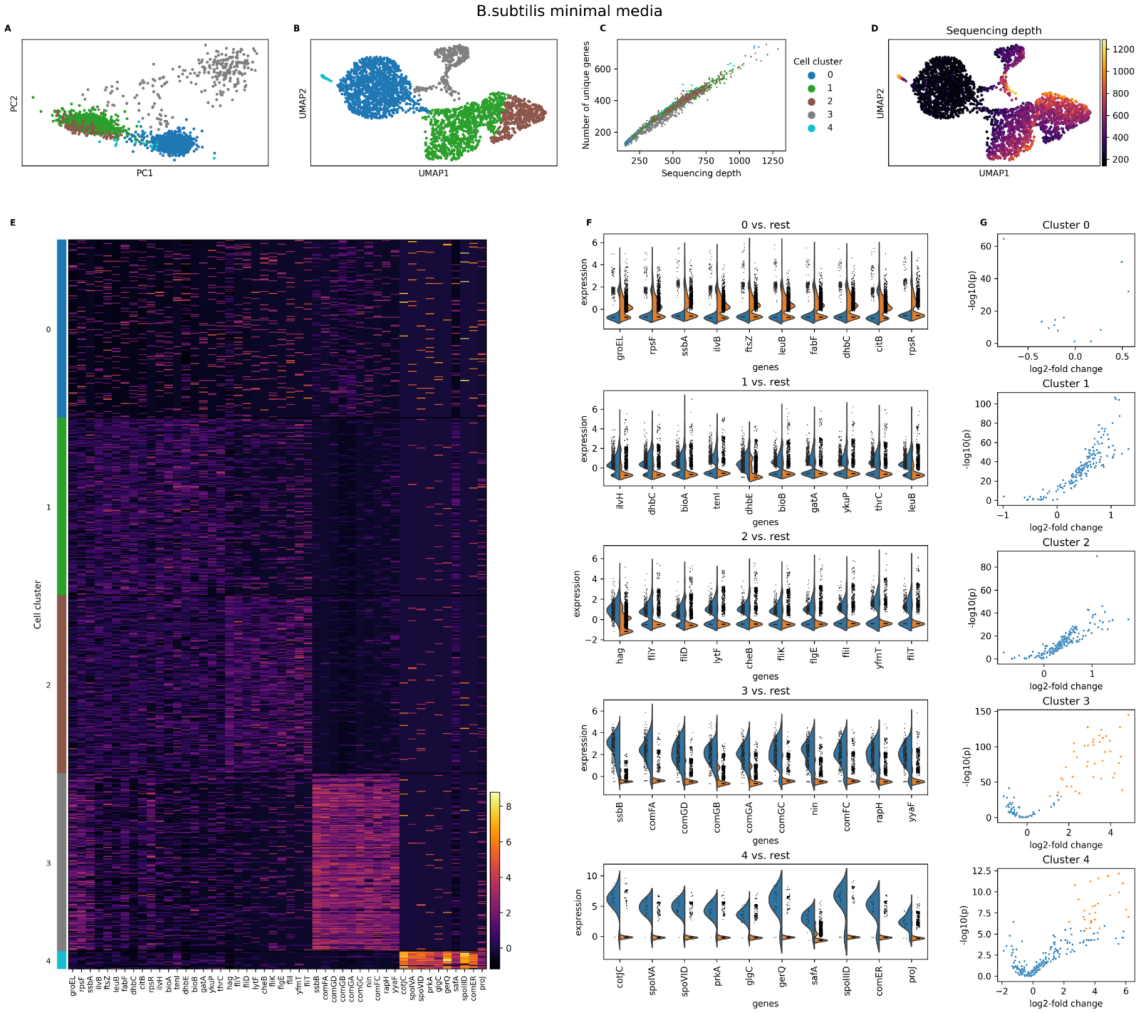


Fig. 3 Transitions between cellular states in *B.subtilis* are pronounced by BacSC. Analysis of the *Bsub_minmed_PB* dataset with BacSC. **(A)** Scatterplot of first two dimensions of PCA embedding with cell type clusters highlighted. **(B)** UMAP plot based on the parameters determined by BacSC, colored by cell cluster. **(C)** Scatterplot of sequencing depth versus number of unique genes per cell, colored by cell cluster. **(D)** Umap plot as in (B), colored by sequencing depth. **(E)** Heatmap of normalized gene expression. For each cluster, the 10 genes with the highest contrast scores are shown. For better visibility of small clusters, at most 200 cells per cluster are shown. **(F)** Violin plots of normalized gene expression for DE tests of each cell type (blue) against the rest of the cell population (orange). For each cluster, the 10 genes with the highest contrast score are shown. **(G)** Volcano plots for DE tests as in (F). The x-axis shows the log-fold change for gene expression, the y-axis shows the $-\log_{10}$-transformed (uncorrected) p-value. Only genes that are expressed in at least 80% of cells in the respective cluster are shown. Orange genes are differentially expressed at the $\alpha = 0.05$-level.

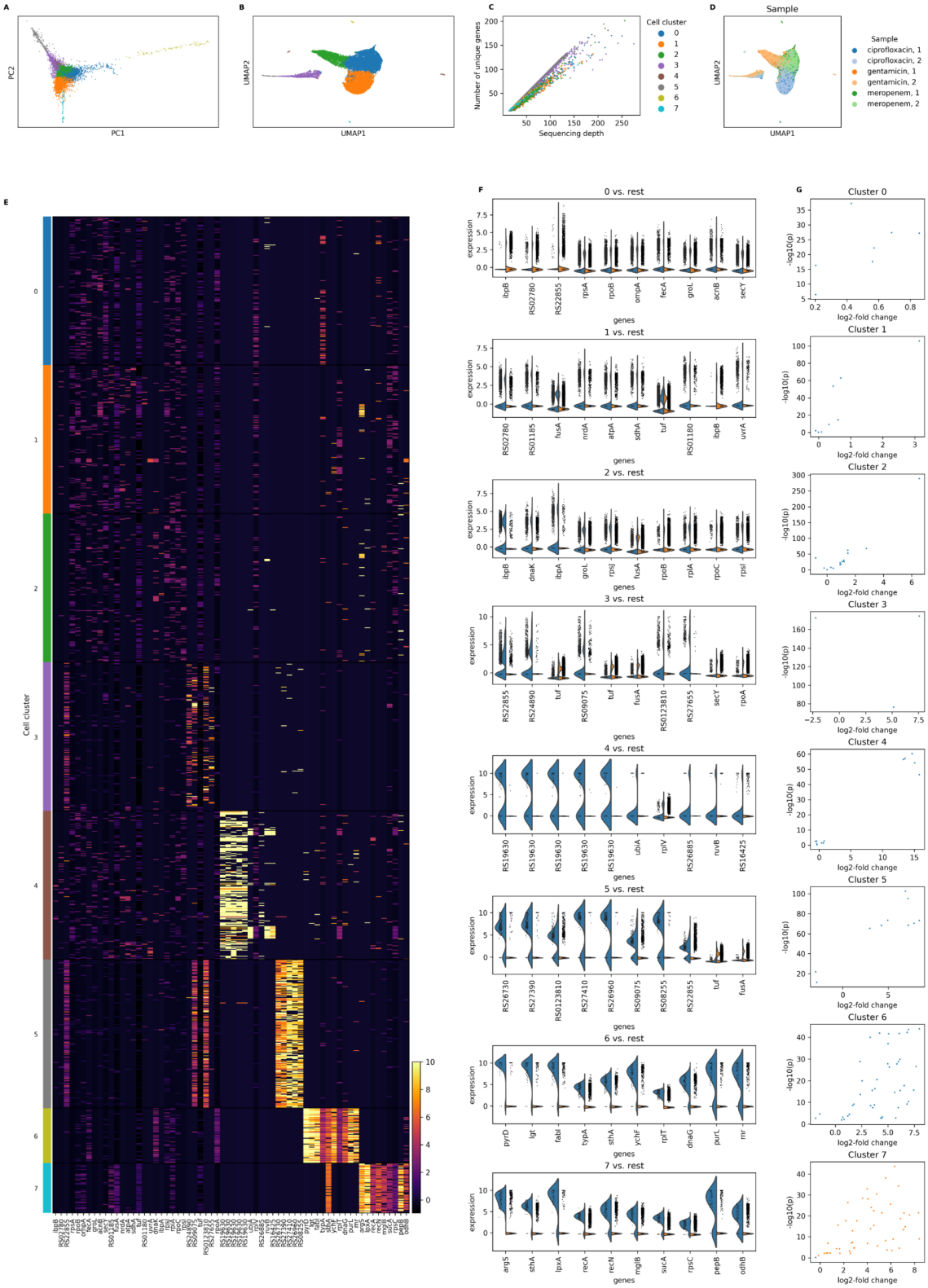


Fig. 4 BacSC shows clear differences in response of *K. pneumoniae* to different antibiotics. Analysis of the *Klebs_antibiotics_BD* dataset with BacSC. **(A)** Scatterplot of first two dimensions of PCA embedding with cell type clusters highlighted. **(B)** UMAP plot based on the parameters determined by BacSC, colored by cell cluster. **(C)** Scatterplot of sequencing depth versus number of unique genes per cell, colored by cell cluster. **(D)** Umap plot as in (B), colored by sample identity (antibiotic and replicate). **(E)** Heatmap of normalized gene expression. For each cluster, the 10 genes with the highest contrast scores are shown. For each cluster, the 10 genes with the highest contrast scores are shown. For better visibility of small clusters, at most 200 cells per cluster are shown. **(F)** Violin plots of normalized gene expression for DE tests of each cell type (blue) against the rest of the cell population (orange). For each cluster, the 10 genes with the highest contrast score are shown. Genes in (E) and (F) are annotated with gene symbols wherever possible, otherwise locus tags are shown. **(G)** Volcano plots for DE tests as in (F). The x-axis shows the log-fold change for gene expression, the y-axis shows the $-\log_{10}$-transformed (uncorrected) p-value. Only genes that are expressed in at least 50% of cells in the respective cluster are shown. Orange genes are differentially expressed at the $\alpha = 0.05$-level.

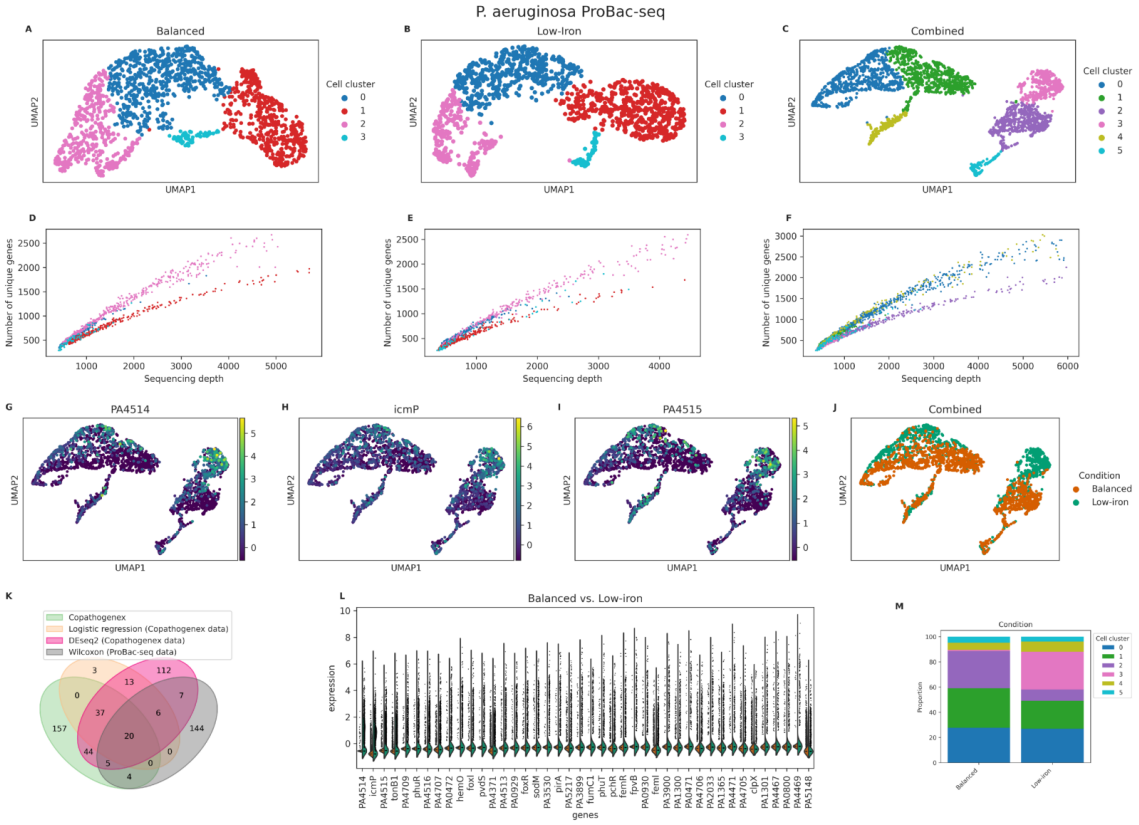


Fig. 5 Processing with BacSC discovers a distinct response of *P. aeruginosa* to a low-iron environment. Analysis of the *Pseudomonas_balanced_PB, Pseudomonas_li_PB* datasets and a combination of both with BacSC. **(A-C)** UMAP plots based on the parameters determined by BacSC, colored by cell cluster for the balanced, low-iron, and combined datasets, respectively. **(D-F)** Scatterplots of sequencing depth versus number of unique genes per cell, colored by cell cluster, for all three datasets. **(G-I)** UMAP plots of the combined dataset, highlighting normalized expression values of the three genes most significantly associated with iron reduction (see subfigure L). **(J)** UMAP plot of the combined dataset, highlighting growth condition (balanced or low-iron) for each cell. **(K)** Venn diagram of differentially expressed genes found in Co-PATHOgenex and ProBac-seq data for Pseudomonas in balanced versus low-iron growth conditions. **(L)** Violin plots of differentially expressed genes in ProBac-seq and Co-PATHOgenex (at least one DE method, balanced vs. low-iron). **(M)** Stacked barplot of cluster proportions for cells from each growth condition.

3 Discussion

The emergence of protocols for scRNA-seq of bacterial populations is about to transform microbiology research by allowing to evaluate the transcriptional profiles of bacteria at an unprecedented combination of scale and resolution. Despite their technological similarity, bacterial scRNA-seq datasets at their current state differ significantly from eukaryotic scRNA-seq data in terms of sparsity and sequencing depth. To facilitate the statistically sound processing of bacterial scRNA-seq data, we present BacSC, a computational pipeline that allows for easy, dataset-specific quality control and automatic variance stabilization, low-dimensional representation, neighborhood embedding, clustering, and differential expression analysis of such data.

By using a variance-stabilizing transform with gene-wise zero imputation parameters [22], BacSC is able to adequately normalize gene expression data with very large amounts of zero entries and low sequencing depth. We show that train-test splitting through data thinning [28, 33] and comparison to negative control data in scDEED [27] provides ways to select suitable parameters for dimensionality reduction, and neighborhood embedding. Furthermore, selecting a clustering resolution through our newly defined gap statistic based on count splitting of the raw expression data reveals biologically distinct subpopulations. To counteract FDR inflation when testing differential gene expression of bacterial cell types, we extend the ClusterDE method [29] to highly disproportionate cluster sizes. Additionally, our copula-based simulation setup adapts the approach from scDesign [47, 56] to bacterial scRNA-seq data. To this end, we add correlation shrinkage [57, 58] and an adjustment for underestimation of small gene-gene correlations.

Overall, BacSC is a highly flexible framework that performs statistical analysis of bacterial scRNA-seq data independent of the underlying sequencing protocol, while avoiding common statistical pitfalls. Through its capabilities for automated parameter selection, BacSC further allows for a set-and-forget approach to bacterial scRNA-seq data processing, greatly simplifying these tasks. We demonstrated this flexibility through application to 13 bacterial scRNA-seq datasets from two protocols across five different species. Despite large differences in size and sequencing depth per cell even after manual quality control, BacSC was able to integrate, cluster, and perform differential expression testing on each dataset without needing any further user intervention.

The detected cell types and their marker genes showed remarkable overlap with the clusters previously found through processing with default or manually selected parameters in multiple datasets [13, 18], confirming the correctness of BacSC's findings. BacSC was further able to better depict dynamics between cellular subpopulations in *B. subtilis* and found new bacterial cell types in *K. pneumoniae*. Analyzing two datasets from *P. aeruginosa* grown in environments with different iron availability, BacSC found similar cell types, highlighting its robustness. After joint processing of both datasets with BacSC, differential expression testing correctly detected various genes related to iron acquisition.

Its modular structure and seamless integration in scanpy [37] allow users to easily apply the entire BacSC pipeline or parts of it to their own data, and perform downstream analysis with other methods provided in the scverse [38]. In our studies, we used these capabilities to test for differential abundance between cell type proportions with scCODA [52].

In addition to the described features, there are multiple areas where further improvements and extensions to BacSC are possible. While we developed and evaluated BacSC with bacterial scRNA-seq data in mind, the techniques used were designed for eukaryotic scRNA-seq analysis. Therefore, BacSC is in principal also suited for this type of data, expanding its application range beyond the usecases shown here.

In its current state, BacSC uses methods that are seen as the baseline in scRNA-seq analysis [25]. While we adapted these techniques here to fit the properties of bacterial scRNA-seq data, there exist a plethora of approaches, each with their own assumptions, that often show improved capabilities on eukaryotic data [59]. Careful evaluation of these methods in the context of bacterial scRNA-seq requires further efforts.

Finally, our improvements on the synthetic data generation algorithm for differential expression testing currently only cover simulation of one homogeneous cell population. An extension to match the capabilities of scDesign2 and scDesign3 [47, 56] in simulating multiple cell types, batches, trajectories, and spatial information is an open challenge.

By eliminating the need to manually select suitable techniques and parameters, BacSC removes sources of errors and allows for more efficient data processing. We therefore believe that BacSC provides an easily applicable framework that facilitates proper statistical analysis of bacterial scRNA-seq data.

Acknowledgments

We thank Sine Lo Svenningsen for providing the *E. coli* strain MAS1081. Furthermore, we thank Petra Hagendorff for her assistance in using the Chromium Controller and Astrid Dröge for her support in preparing sequencing libraries.

C.L.M. acknowledges core funding from the Institute of Computational Biology, Helmholtz Zentrum München. C.L.M. and S.H. received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the framework of the Priority Program SPP2389 "Emergent functions of bacterial multicellularity" (HA 3299/9-1, AOBJ: 687646). Furthermore, S.H. received funding under Germany's Excellence Strategy – EXC 2155 "RESIST" – Project ID 390874280, from the Novo Nordisk Foundation (NNF 18OC0033946), from the SFB/TRR-298-SIIRI – Project-ID 426335750 and the Ministry of Science and Culture of Lower Saxony (Niedersächsisches Ministerium für Wissenschaft und Kultur) BacData, ZN3428. R.O.A. and C.L.M. were funded by the StressRegNet consortium within the Bavarian research network bayresq.net funded through the Bavarian State Ministry of Science and Arts, Germany.

Author contributions

J.O. and C.L.M. designed the structure and individual steps of the BacSC pipeline, and conceived improvements to existing methods. T.K., J.G.T. and S.H. generated data containing *P. aeruginosa* and *E. coli* with ProBac-seq, A.Z.R. provided all datasets from *B. subtilis*. J.O. implemented the pipeline and conducted all applications and tests. J.O., T.K., R.O.A., J.G.T., and A.Z.R. analyzed the results from BacSC in a biological context, R.O.A. further performed analysis of the Co-PATHOgenex data. J.O. wrote the manuscript with help from all other authors. All authors read and approved the manuscript.

Declaration of Interests

The authors declare no competing interests.

Supplemental information

- Supplemental pdf:
  - Additional dataset analysis
  - Supplemental figures B1-B5, C6-C17, D18-D31
  - Supplemental tables E1-E17
- Pa_probes.xslx: Probes used in ProBac-seq of *P. aeruginosa*

STAR Methods

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Johannes Ostner (johannes.ostner@stat.uni-muenchen.de).

Materials Availability

Materials generated in this study are freely available at public repositories (see key resources table) or by contacting the lead contact.

Data and Code Availability

Single-cell RNA-seq data have been deposited at GEO and are publicly available as of the date of publication. Accession numbers are listed in the key resources table. Intermediate datasets have been deposited at zenodo and are publicly available as of the date of publication. DOIs are listed in the key resources table. This paper analyzes existing, publicly available data. The accession numbers for these datasets are listed in the key resources table. All original code has been deposited at GitHub (https://github.com/bio-datascience/BacSC) and is publicly available as of the date of publication. DOIs are listed in the key resources table. (additional citations in the key resources table: [60–62])

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

For validating the performance of BacSC, we analyzed previously published scRNA-seq datasets. For ProBac-seq data analysis, we used the *Bsub_minmed_PB* dataset from the original publication (GEO: GSE223752) [18]. For BacDrop data analysis, we selected seven datasets provided in the original publication [13], and used the read count matrices published by the authors (GEO: GSE180237). The *Klebs_BIDMC35_BD*, *Klebs_4species_BD*, *Ecoli_4species_BD*, *Efaecium_4species_BD*, and *Pseudomonas_4species_BD* datasets were used as provided. For the *Klebs_untreated_BD*, and *Klebs_antibiotics_BD* datasets, we concatenated the count matrices from multiple samples before analysis wth BacSC.

Furthermore, in this study we generated additional datasets using ProBac-seq, encompassing two experiments on *Bacillus subtilis*, two samples of *Pseudomonas aeruginosa*, as well as one sample of *Escherichia coli*.

ProBac-seq of B. subtilis

For the *Bsub_damage_PB* dataset, cells were grown to mid-log phase in spizizen's minimal media (SMM) and Mitomycin C (MMC, $0.5\mu g/ml$ final concentration) was added to wildtype *B.subtilis* (strain 168) as reported by [63]. The *Bsub_MPA_PB* data contains *B.subtilis* cells grown in SMM as described by [64, 65] to mid-log phase and challenged with Mycophenolic acid (MPA, $40\mu g/ml$ final concentration).

ProBac-seq of E. coli and P. aeruginosa

For the samples *Ecoli_balanced_PB*, *Pseudomonas_balanced_PB* and *Pseudomonas_li_PB* MOPS (morpholinepropanesulfonic acid) minimal medium (supplemented with 100 ng/µl thiamine) with 0.2 % glucose as the sole carbon source was used [66]. To induce a mild iron limitation on *Pseudomonas_li_PB*, the $FeSO_4$ concentration was lowered to 0.5 µM instead of the regular 10 µM. Single colonies of *E. coli* MAS1081 [67, 68] and PAO1 were used to inoculate precultures with regular MOPS and were grown for 11-12 hours at 37°C with shaking at 180 rpm. After washing, main cultures in MOPS with normal iron or reduced iron content were inoculated at an OD_{600} of 0.00002 and grown for 10-14 generations. Bacteria were harvested in balanced growth conditions in early exponential phase (OD_{600} of 0.2-0.3).

METHOD DETAILS

ProBac-seq of B. subtilis

For all *B. subtilis* datasets, ProBac-seq was performed as described in the original method [18, 36].

ProBac-seq of E. coli and P. aeruginosa

Further sample preparation for ProBac-seq was performed as previously described [18, 36], with slight modifications. In brief, 1 ml of each culture was used for fixation with 1 % formaldehyde for 30 min at room temperature. To increase the cell yield, all centrifugation steps were carried out at 7,000 x g for up to 5 min. Overnight storage in MAAM (4:1 V:V dilution of methanol to acetic acid) was omitted. All further steps were performed according to the protocol of the original method [18, 36]. PAO1-specific probes were designed and generated as previously described without additional UMI extension. The single-cell sequencing libraries were quality-checked and sequenced by the GMAK sequencing facility (HZI, Braunschweig, Germany) on a NovaSeq SP flow cell (100 cycles, 28-10-10-90) resulting in up to 170 million reads per sample. Raw fastq files were processed with CellRanger v7.1.0 [69] with the option *–expect-cells 10000*.

QUANTIFICATION AND STATISTICAL ANALYSIS

This section describes statistical details for the individual steps in the BacSC pipeline. Statistical details and results from application of the BacSC pipeline to all datasets described in table 1 can be found in supplementary figures D18-D31 and supplementary tables E1-E17.

Processing starts with a raw counts matrix $X_0 \in \mathbb{N}_0^{n_0 \times p_0}$, which contains read counts of p_0 genes for n_0 droplets.

Quality control

For datasets generated with ProBac-seq, multiple probe reads for each gene are available. As described in the original publications [18, 36], we aggregated the probes by max-pooling. Furthermore, most datasets from ProBac-seq were already quality-controlled in CellRanger [69] and therefore needed less additional filtering. For all ProBac-seq datasets, we chose a minimum sequencing depth cutoff of 100. For data from BacDrop, we used the minimum sequencing depth cutoff of 15, as provided in the original publication [13]. For the three largest datasets (*Klebs_untreated_BD, Klebs_antibiotics_BD, Klebs_BIDMC35_BD*), we also selected 2,500 highly variable genes after variance stabilization. BacSC further removes genes that were expressed in only a single cell, as variance stabilization for these genes is not possible. In contrast to eukaryotic scRNA-seq datasets, removal of mitochondrial genes is not required for bacterial scRNA-seq, as bacteria do not contain mitochondria. Still, other highly abundant types of RNA, such as rRNA and tmRNA, can be removed at this point. For the analysis presented here, we did not perform any removal of features beyond the preprocessing in CellRanger [69] for ProBac-seq or UMI-tools [70] for BacDrop.

Further outliers are detected by filtering cells based on median absolute deviations (MAD) of their log-transformed total counts and number of expressed genes [30]: $MAD(S) = median_{i=1}^{n}(|\log(S_i) - median(\log(S))|)$ where S is either the vector of sequencing depths $\sum_{j=1}^{p_0} X_{\cdot,j}$ or number of expressed genes over all cells. A cell is considered an outlier if for either of the two metrics, $|S_i - median(S)| > nmads * MAD(S)$, where $nmads$ is the factor defined in table E2.

Table E2 gives an overview over the filtering parameters chosen for each dataset. After filtering, X_0 is reduced to a matrix $X \in \mathbb{N}_0^{n \times p}$ of p genes and n cells.

Variance stabilization

For variance-stabilizing transformation (VST) of the filtered read counts, we follow the results from [22]. Assuming potential overdispersion of the count distribution, we use an approximation to the ideal VST determined by the delta method, a log-transformation in combination with common-sum scaling of the counts:

$$\tilde{X}_{i,j} = \log\left(\frac{X_{i,j}}{m_i} + \nu\right) \tag{1}$$

where $m_i = \frac{\sum_{j=1}^{p} X_{i,j}}{median_{k=1}^{n}(\sum_{j=1}^{p} X_{k,j})}$ scales each cell's counts to the median value of all sequencing depths. We chose the median sequencing depth as a scaling factor to gain robustness to outliers in sequencing depth.

Adding a pseudocount ν before log-transformation is necessary to handle zero entries in X. As described in [22], we set $\nu_j = \frac{\theta_j}{4}$ for each gene $j = 1 \dots p$, where θ_j denotes the gene's overdispersion factor. Calculating this overdispersion factor is not straightforward for genes with very low numbers of expressed genes, as the relation $\theta_j = \frac{mean(X_{\cdot,j})^2}{Var(X_{\cdot,j} - mean(X_{\cdot,j}))}$ becomes very sensitive to single entries in X. Instead, we make use of the gene overdispersion estimates provided by sctransform [41], which jointly

models all genes, and thus produces more robust estimates of θ_j. To this end, we apply sctransform to the count matrix X, extract the overdispersion estimates, and use them in equation 1.

After VST, we scale each gene individually to zero mean and unit variance by applying scanpy's *scale* function [37], clipping large values at 10. This results in a normalized gene expression matrix $Y \in \mathbb{R}^{n \times p}$.

Dimension reduction

The selection of the best embedding dimensionality k_{opt} through data thinning was described for Poisson-distributed data in [31]. There, data thinning [33] is used to split the raw count data X into two $n \times p$-dimensional datasets X^{train} and X^{test} by a random binomial split on each individual entry in X. The resulting train and test matrices are then both Poisson-distributed again. Because eukaryotic single-cell data is typically assumed to follow a Negative Binomial (NB) distribution for each gene, [28] extended the data-thinning approach to NB-distributed data. However, the lower read counts in bacterial scRNA-seq suggest that the data might follow a linear instead of a quadratic mean-variance pattern and are therefore Poisson-distributed.

To determine the distributional assumption for count splitting, we first calculate the mean μ_j and variances σ_j^2 of $X_{\cdot,j}$ for each gene $j = 1 \dots p$. We then compare Pearson correlation coefficients r of a linear and a quadratic relation between μ and σ^2. If $r_{quadratic} > r_{linear}$, we assume X to be Negative Binomial distributed, otherwise it is Poisson-distributed. The raw data distribution for each dataset is shown in Table E2.

Depending on the chosen data distribution, X is split into two datasets by Poisson or NB count splitting (Figure B1A, B). In both cases, we set the split ratio $\epsilon = 0.5$ to ensure an even split between train and test data and maximize the probability of obtaining a nonzero entry in train and test data if $X_{i,j} > 1$. We then determine all genes or cells that have only one nonzero entry in X_{train} or X_{test}, and remove them from both data splits. In line with [31], we apply the VST described in section 3 to both X_{train} and X_{test}, using the θ parameters determined on the whole data to speed up computation, and obtain transformed matrices Y_{train} and Y_{test}.

To determine k_{opt}, we perform a singular value decomposition (SVD) $Y_{train} = U\Sigma V^T$ on the training data. For each $k = 1 \dots 20$, we then calculate the reconstruction loss as sum of squared differences between the test data and the k-dimensional approximation of the SVD of the train data (Figure B1C):

$$\begin{aligned} L_k &= ||Y_{test} - U_{\cdot,1:k}\Sigma_{1:k,1:k}V^T_{\cdot,1:k}||^2_F \\ k_{opt} &= \underset{k=1\dots20}{\arg\min}\, L(k) \end{aligned} \tag{2}$$

Data visualization

BacSC selects the latent parameters $n_{neighbors}$ and min_{dist} for constructing a UMAP embedding of the data through scDEED [27]. For every combination of $n_{neighbors}$ (the number of neighbors for each cell in the neighborhood graph) and min_{dist} (the effective minimum distance between points), scDEED defines a reliability score for each cell as the Pearson correlation between the euclidean distances to the 50% closest cells in PCA space and the euclidean distances to these cells after UMAP embedding. To obtain a baseline distribution, another set of reliability scores is calculated on a permuted dataset where each gene's expression values are shuffled. scDEED then classifies the embedding of cells in the original dataset as "trustworthy", "undefined", or "dubious" based on the 95% and 5% quantiles of the distribution of reliability scores in the permuted data (Figure B1D). Finally, the parameter combination with the smallest number of dubiously embedded cells is selected (Figure B1E, F).

As scDEED is only available in R, the BacSC pipeline includes a Python implementation of the method. For every dataset, we considered all pairwise combinations of parameters: $n_{neighbors}$: $(10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 150, 200, 250)$; min_{dist} : $(0.05, 0.1, 0.3, 0.5, 0.7)$.

Clustering

The resolution parameters in Louvain and Leiden clustering are essential for defining the granularity of the resulting partition [44, 71]. Both algorithms aim to optimize the modularity or a similar metric of a partition on the neighborhood graph defined during UMAP generation:

$$Modularity = \frac{1}{2m}\sum_c (e_c - \gamma\frac{K_c^2}{2m}) \tag{3}$$

where m is the total number of edges in the neighborhood graph, e_c is the number of edges within cluster c, and K_c is the sum of degrees over all nodes in cluster c, and γ is the resolution parameter. Generally, a higher resolution parameter will lead to a more fine-grained clustering. While both algorithms effectively approximate an optimal clustering for a given value of γ, the choice of a "good" resolution parameter is highly dependent on the structure and biological source of the data at hand [32, 72]. Larger datasets or datasets from more complex communities generally contain more subclusters and thus warrant a larger value of γ to detect all relevant subpopulations. On the other hand, choosing the resolution too large will result in non-robust clusterings that are highly sensitive to small perturbations of the data [73]. Furthermore, cluster assignments and the number of subpopulations are not monotonic in γ, complicating the evaluation of clustering quality [32]. In BacSC, we aim to automatically find a resolution parameter that results in an informative, but stable clustering of the cells.

To this end, we adapt the idea from [28] and use the train and test datasets obtained through count splitting for clustering evaluation. Starting with the variance-stabilized train and test data from dimensionality reduction, we generate the neighborhood graph for both datasets with the k and $n_{neighbors}$ parameters determined earlier. For each value of γ in a set of possible resolutions, we then perform Leiden or Louvain clustering on the training data, resulting in a cluster assignment c_{train}. Since training and test data contain the same cells, we can now obtain a measure for the robustness of the clustering by calculating the modularity (3) for c_{train} on the neighborhood graph of the test data (Figure B2A). We denote this value with M_{test}. Since modularity generally decreases with the number of clusters, we cannot select the value of γ for which M_{test} is maximal. Instead, we need to compare the test data resolution to a baseline score for each resolution value. Therefore, we generate a random cluster assignment on the test data by permuting the labels from c_{train} and calculate M_{random}, the modularity of the random clustering on the neighborhood graph of the test data. Finally, we select the resolution where the gap statistic between test modularity and random modularity is maximal (Figure B2B, C):

$$res_{opt} = \arg\max_{\gamma}(M_{test} - M_{ramdom}) \tag{4}$$

and perform a clustering with res_{opt} on the full dataset to obtain cell type clusters (Figure B2D). For processing the datasets in this manuscript, we used the Leiden algorithm and modularity score and tested possible resolutions $\gamma = (0.01, 0.03, 0.05, \ldots 0.49)$. The same procedure is however also applicable to Louvain clustering or other measures, e.g. the Constant Potts model [74].

Even though the resolution value determined by maximizing our gap statistic provides improvement over random cluster assignment while being robust to small data perturbations, it is by no means the only "correct" resolution value. For some datasets, more fine-grained clusterings can give further insights into subpopulations of the data. Rather, res_{opt} may serve as a baseline clustering resolution that gives an adequate first insight into the data.

Differential expression testing

Identifying genes with characteristic expression for cell clusters defined by the same gene expression values is an instance of reusing information, or "double dipping" [46], and controlling the false discovery rate under such conditions is essential to achieve adequate results. The ClusterDE method [29] provides FDR control for DE testing of cell types in eukaryotic scRNA-seq by contrasting the p-values of interest with p-values calculated on a synthetically generated negative control dataset. In BacSC, we implement a modified version of the algorithm that takes the characteristics of bacterial single-cell data into account and allows for testing of highly disproportionate cell populations. The following description assumes a DE test of cell type C with n_C cells against the union of all other cell types, containing $n_{\bar{C}} = n - n_C$ cells (Figure B3A). Tests of differential gene expression between two cell types are possible in the same manner, but the data needs to be subsetted to the clusters of interest first.

ClusterDE first generates negative control data with the same marginal gene distributions and gene-gene correlations as the original data, but no intrinsic cluster structure. This synthetic data generation is done with scDesign2 [47] or scDesign3 [56], which both use a Gaussian copula approach to generate synthetic scRNA-seq data. To account for the high sparsity and low sequencing depth of bacterial scRNA-seq data, we adapted the approach from scDesign2 in BacSC. In a first step, the marginal distribution of raw counts is determined for every gene j. As in scDesign2, we consider four possible distributions - Poisson (Poi), zero-inflated Poisson (ZIP), Negative Binomial (NB), and zero-inflated Negative Binomial (ZINB). If the gene's empirical variance σ_j^2 is larger than its empirical mean μ_j, we determine the gene to be NB- or ZINB-distributed, otherwise its distribution is Poi or ZIP. We then fit the Poisson or NB distribution with and without zero-inflation to $X_{\cdot,j}$ through maximum likelihood estimation via BFGS,

as implemented in the *statsmodels* package [75]. Because of the large number of zeros, we experienced frequent convergence problems with NB estimation. To counteract this, we set the initial mean and dispersion parameters for both NB and ZINB to the mean and dispersion of all nonzero entries in $X_{\cdot,j}$, and the initial zero inflation in the ZINB model to the proportion of zeros in $X_{\cdot,j}$. If both the NB and ZINB models still do not converge, we instead use the estimates from the NB model with default starting parameters, regardless of convergence. We then perform a likelihood-ratio test between the log-likelihoods of the zero-inflated and regular model. If the null hypothesis of no difference in log-likelihood between both models is rejected at the $\alpha = 0.05$ level, we model the gene with zero-inflation, otherwise we use the non-zero-inflated estimate. Denote the chosen distribution for gene j with its estimated parameters as as $D_j(\phi_j)$

As in scDesign2, we now transform the discrete counts for each gene to continuous quantiles through a uniform approximation with the corresponding cumulative distribution function (CDF) $\hat{D}_j(\phi_j)$:

$$U_{\cdot,j} = V_j \hat{D}_j(X_{\cdot,j}, \phi_j) + (1 - V_j)\hat{D}_j(X_{\cdot,j} + 1, \phi_j) \tag{5}$$

with $V_j \sim Uniform(0,1)^n$. We then transform these quantiles by the inverse CDF (denoted Φ^{-1}) of a standard normal distribution and calculate their empirical correlation matrix $R \in \mathbb{R}^{p \times p}$.

Contrary to eukaryotic scRNA-seq, where current datasets contain many more cells than genes, most of our bacterial scRNA-seq data is underdetermined, with $n < p$ (Table E1). Therefore, the entries of the empirical covariance matrix must be shrunk to obtain a good estimate for R [57, 58]. To this end, we use a Python reimplementation of the covariance shrinkage proposed in [76].

The uniform approximation 5 in the copula transformation is necessary to allow the use of Gaussian copula for discrete count data, but shifts the count matrix by an average of 0.5. Since bacterial scRNA-seq data contains mostly zero or very small entries, this leads to considerably lower gene-gene correlations and gene variances in the generated data. To counteract this, we introduce a scaling factor δ on off-diagonal entries of R where the absolute absolute value of the original data's gene-gene correlation S is larger than 0.1:

$$\hat{R}_{i,j}(\delta) = \begin{cases} \delta R_{i,j}, & \text{if } |S_{i,j}| > 0.1 \\ R_{i,j}, & \text{otherwise} \end{cases} \tag{6}$$

The scaled correlation matrix $\hat{R}(\delta)$ is not guaranteed to be positive definite though. To obtain a positive definite matrix $\tilde{R}(\delta)$ that is close to $\hat{R}(\delta)$, we calculate the eigendecomposition (λ, v) of $\hat{R}(\delta)$, increase all eigenvalues by $-\lambda_{min} + 10^{-12}$ if the smallest eigenvalue λ_{min} is negative, and set $\tilde{R}(\delta) = v\ diag(\tilde{\lambda})\ v^{-1}$ with the shifted eigenvalues $\tilde{\lambda}$. We then determine the ideal δ through a golden ratio optimizer [77] with initial bracket $(1, 2)$ that minimizes the sum of squared differences between the scaled entries of $\tilde{R}(\delta)$ and S:

$$\delta^* = \arg\min_{\delta} \sum_{(i,j):|S_{i,j}|>0.1} (S_{i,j} - \tilde{R}(\delta)_{i,j})^2 \tag{7}$$

Scaling of the entries in R will slightly overestimate the gene means of the generated data (Figure B3B), but gives better results for large gene variances and gene-gene correlations (Figure B3C, D). To simulate synthetic null data with n' samples and no apparent cluster structure, we generate n' samples $\hat{Z}$ from a $Normal(0, \tilde{R}(\delta^*)$ distribution, and transform them back into the original space by the standard normal CDF and the inverse CDF of $D_j(\phi_j)$:

$$\hat{X}_{\cdot,j} = \hat{D}_j^{-1}(\Phi(\hat{Z}_{\cdot,j})) \in \mathbb{N}_0^{n' \times p} \tag{8}$$

Using this procedure, we can obtain a synthetic null dataset with marginal distributions and gene-gene correlations similar to the target data, but no cluster structure. To allow for generation of negative control data that has the same numbers of cells in both groups as the original data, we set $n' = 2n$ and subset $\hat{X}$ after processing. Analogous to ClusterDE, we process the synthetic null data in the same way as the original data. We use the same parameters for dimension reduction and neighborhood embedding as determined for the target data, but re-run sctransform on the null data to get new estimates for the gene-wise overdispersion θ. By finding a suitable resolution for the Leiden algorithm, we cluster $\hat{X}$ into exactly two parts, and randomly draw n_C and $n_{\bar{C}}$ cells from both clusters, respectively (Figure B3E).

FDR control in ClusterDE and BacSC is performed through contrast scores and the Clipper method [35]. We first obtain two sets of n p-values by performing the same DE test (e.g. Wilcoxon rank-sum) on the original data and on the drawn subset of the synthetic null data (Figure B3F, G). Next, we calculate the contrast score

$$\Gamma_i = (-log_{10}(p_{data,i}) - (-log_{10}(p_{null,i}))) \tag{9}$$

for each pair of p-values. Given a FDR level α, Clipper then finds a threshold T on the contrast scores

$$T = \min \left\{ 0 < t < \max(\Gamma) : \frac{|\{i : \Gamma_i \leq -t\}| + 1}{|\{i : \Gamma_i \geq t\}| \vee 1} \leq \alpha \right\} \tag{10}$$

For genes with $\Gamma_i > T$, the expected FDR is less than α [34] and we denote them as differentially expressed (Figure B3H).

While differential expression testing with contrast scores is not computationally intensive, the generation of synthetic null data does require some computational power. Fortunately, a series of tests of each cell type's gene expression against the union of all other cell types only requires generation of the synthetic null data once, as the same set of cells is included in every test and therefore marginal gene distributions and correlations are identical. Only the selection of n_C and $n_{\bar{C}}$ cells from $\hat{X}$ and subsequent steps have to be performed individually for each cell type.

References

[1] Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., B Tuch, B., Siddiqui, A., Lao, K., Azim Surani, M.: mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods **6**(5), 377–382 (2009) https://doi.org/10.1038/nmeth.1315

[2] Haber, A.L., Biton, M., Rogel, N., Herbst, R.H., Shekhar, K., Smillie, C., Burgin, G., Delorey, T.M., Howitt, M.R., Katz, Y., Tirosh, I., Beyaz, S., Dionne, D., Zhang, M., Raychowdhury, R., Garrett, W.S., Rozenblatt-Rosen, O., Shi, H.N., Yilmaz, O., Xavier, R.J., Regev, A.: A single-cell survey of the small intestinal epithelium. Nature **551**(7680), 333–339 (2017) https://doi.org/10.1038/nature24489

[3] Montoro, D.T., Haber, A.L., Biton, M., Vinarsky, V., Lin, B., Birket, S.E., Yuan, F., Chen, S., Leung, H.M., Villoria, J., Rogel, N., Burgin, G., Tsankov, A.M., Waghray, A., Slyper, M., Waldman, J., Nguyen, L., Dionne, D., Rozenblatt-Rosen, O., Tata, P.R., Mou, H., Shivaraju, M., Bihler, H., Mense, M., Tearney, G.J., Rowe, S.M., Engelhardt, J.F., Regev, A., Rajagopal, J.: A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. Nature **560**(7718), 319–324 (2018) https://doi.org/10.1038/s41586-018-0393-7

[4] Han, A., Glanville, J., Hansmann, L., Davis, M.M.: Linking t-cell receptor sequence to functional phenotype at the single-cell level. Nat. Biotechnol. **32**(7), 684–692 (2014) https://doi.org/10.1038/nbt.2938

[5] Roux, A.E., Yuan, H., Podshivalova, K., Hendrickson, D., Kerr, R., Kenyon, C., Kelley, D.: Individual cell types in c. elegans age differently and activate distinct cell-protective responses. Cell Rep. **42**(8), 112902 (2023) https://doi.org/10.1016/j.celrep.2023.112902

[6] McFarland, J.M., Paolella, B.R., Warren, A., Geiger-Schuller, K., Shibue, T., Rothberg, M., Kuksenko, O., Colgan, W.N., Jones, A., Chambers, E., Dionne, D., Bender, S., Wolpin, B.M., Ghandi, M., Tirosh, I., Rozenblatt-Rosen, O., Roth, J.A., Golub, T.R., Regev, A., Aguirre, A.J., Vazquez, F., Tsherniak, A.: Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. Nat. Commun. **11**(1), 4296 (2020) https://doi.org/10.1038/s41467-020-17440-w

[7] Walls, A.W., Rosenthal, A.Z.: Bacterial phenotypic heterogeneity through the lens of single-cell RNA sequencing. Transcription **15**(1-2), 48–62 (2024) https://doi.org/10.1080/21541264.2024.2334110

[8] Gu, J., Lin, Y., Wang, Z., Pan, Q., Cai, G., He, Q., Xu, X., Cai, X.: Campylobacter jejuni cytolethal distending toxin induces GSDME-Dependent pyroptosis in colonic epithelial cells. Front. Cell. Infect. Microbiol. **12** (2022) https://doi.org/10.3389/fcimb.2022.853204

[9] Cerny, O., Godlee, C., Lobato-Márquez, D.: Editorial: Single cell analysis of bacteria-host interaction. Front. Cell. Infect. Microbiol. **13** (2023) https://doi.org/10.3389/fcimb.2023.1196905

[10] Jia, M., Zhu, S., Xue, M.-Y., Chen, H., Xu, J., Song, M., Tang, Y., Liu, X., Tao, Y., Zhang, T., Liu, J.-X., Wang, Y., Sun, H.-Z.: Single-cell transcriptomics across 2,534 microbial species reveals functional heterogeneity in the rumen microbiome. Nat Microbiol (2024) https://doi.org/10.1038/s41564-024-01723-9

[11] Lötstedt, B., Stražar, M., Xavier, R., Regev, A., Vickovic, S.: Spatial host–microbiome sequencing reveals niches in the mouse gut. Nat. Biotechnol., 1–10 (2023) https://doi.org/10.1038/s41587-023-01988-1

[12] Brennan, M.A., Rosenthal, A.Z.: Single-Cell RNA sequencing elucidates the structure and organization of microbial communities. Front. Microbiol. **12**, 713128 (2021) https://doi.org/10.3389/fmicb.2021.713128

[13] Ma, P., Amemiya, H.M., He, L.L., Gandhi, S.J., Nicol, R., Bhattacharyya, R.P., Smillie, C.S., Hung, D.T.: Bacterial droplet-based single-cell RNA-seq reveals antibiotic-associated heterogeneous cellular states. Cell (2023) https://doi.org/10.1016/j.cell.2023.01.002

[14] Kuchina, A., Brettner, L.M., Paleologu, L., Roco, C.M., Rosenberg, A.B., Carignano, A., Kibler, R., Hirano, M., DePaolo, R.W., Seelig, G.: Microbial single-cell RNA sequencing by split-pool barcoding. Science **371**(6531) (2021) https://doi.org/10.1126/science.aba5257

[15] Jenniches, L., Michaux, C., Popella, L., Reichardt, S., Vogel, J., Westermann, A.J., Barquist, L.: Improved RNA stability estimation through bayesian modeling reveals most *Salmonella* transcripts have subminute half-lives. Proc. Natl. Acad. Sci. U. S. A. **121**(14), 2308814121 (2024) https://doi.org/10.1073/pnas.2308814121

[16] Wang, B., Lin, A.E., Yuan, J., Novak, K.E., Koch, M.D., Wingreen, N.S., Adamson, B., Gitai, Z.: Single-cell massively-parallel multiplexed microbial sequencing (m3-seq) identifies rare bacterial populations and profiles phage infection. Nat Microbiol **8**(10), 1846–1862 (2023) https://doi.org/10.1038/s41564-023-01462-3

[17] Homberger, C., Hayward, R.J., Barquist, L., Vogel, J.: Improved bacterial Single-Cell RNA-Seq through automated MATQ-Seq and Cas9-Based removal of rRNA reads. MBio **14**(2), 0355722 (2023) https://doi.org/10.1128/mbio.03557-22

[18] McNulty, R., Sritharan, D., Pahng, S.H., Meisch, J.P., Liu, S., Brennan, M.A., Saxer, G., Hormoz, S., Rosenthal, A.Z.: Probe-based bacterial single-cell RNA sequencing predicts toxin regulation. Nat Microbiol **8**(5), 934–945 (2023) https://doi.org/10.1038/s41564-023-01348-4

[19] Blattman, S.B., Jiang, W., Oikonomou, P., Tavazoie, S.: Prokaryotic single-cell RNA sequencing by in situ combinatorial indexing. Nat Microbiol **5**(10), 1192–1201 (2020) https://doi.org/10.1038/s41564-020-0729-6

[20] Kharchenko, P.V.: The triumphs and limitations of computational methods for scRNA-seq. Nat. Methods **18**(7), 723–732 (2021) https://doi.org/10.1038/s41592-021-01171-x

[21] Chari, T., Pachter, L.: The specious art of single-cell genomics. PLoS Comput. Biol. **19**(8), 1011288 (2023) https://doi.org/10.1371/journal.pcbi.1011288

[22] Ahlmann-Eltze, C., Huber, W.: Comparison of transformations for single-cell RNA-seq data. Nat. Methods (2023) https://doi.org/10.1038/s41592-023-01814-1

[23] Luecken, M.D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colomé-Tatché, M., Theis, F.J.: Benchmarking atlas-level data integration in single-cell genomics. Nat. Methods **19**(1), 41–50 (2021) https://doi.org/10.1038/s41592-021-01336-8

[24] Luecken, M.D., Theis, F.J.: Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. **15**(6), 8746 (2019) https://doi.org/10.15252/msb.20188746

[25] Heumos, L., Schaar, A.C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M.D., Strobl, D.C., Henao, J., Curion, F., Single-cell Best Practices Consortium, Schiller, H.B., Theis, F.J.: Best practices for single-cell analysis across modalities. Nat. Rev. Genet., 1–23 (2023) https://doi.org/10.1038/s41576-023-00586-w

[26] Andrews, T.S., Kiselev, V.Y., McCarthy, D., Hemberg, M.: Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. Nat. Protoc. **16**(1), 1–9 (2020) https://doi.org/10.1038/s41596-020-00409-w

[27] Xia, L., Lee, C., Li, J.J.: scDEED: a statistical method for detecting dubious 2D single-cell embeddings (2023). https://doi.org/10.1101/2023.04.21.537839 . https://www.biorxiv.org/content/10.1101/2023.04.21.537839v1

[28] Neufeld, A., Popp, J., Gao, L.L., Battle, A., Witten, D.: Negative binomial count splitting for single-cell RNA sequencing data (2023) arXiv:2307.12985 [stat.ME]

[29] Song, D., Li, K., Ge, X., Li, J.J.: ClusterDE: a post-clustering differential expression (DE) method

robust to false-positive inflation caused by double dipping. bioRxiv (2023) https://doi.org/10.1101/2023.07.21.550107

[30] Germain, P.-L., Sonrel, A., Robinson, M.D.: pipecomp, a general framework for the evaluation of computational pipelines, reveals performant single cell RNA-seq preprocessing tools. Genome Biol. **21**(1), 227 (2020) https://doi.org/10.1186/s13059-020-02136-7

[31] Neufeld, A., Dharamshi, A., Gao, L.L., Witten, D.: Data thinning for convolution-closed distributions (2023) arXiv:2301.07276 [stat.ME]

[32] Grabski, I.N., Street, K., Irizarry, R.A.: Significance analysis for clustering with single-cell RNA-sequencing data. Nat. Methods **20**(8), 1196–1202 (2023) https://doi.org/10.1038/s41592-023-01933-9

[33] Dharamshi, A., Neufeld, A., Motwani, K., Gao, L.L., Witten, D., Bien, J.: Generalized data thinning using sufficient statistics (2023) arXiv:2303.12931 [stat.ME]

[34] Barber, R.F., Candès, E.J.: Controlling the false discovery rate via knockoffs. aos **43**(5), 2055–2085 (2015) https://doi.org/10.1214/15-AOS1337

[35] Ge, X., Chen, Y.E., Song, D., McDermott, M., Woyshner, K., Manousopoulou, A., Wang, N., Li, W., Wang, L.D., Li, J.J.: Clipper: p-value-free FDR control on high-throughput data from two conditions. Genome Biol. **22**(1), 288 (2021) https://doi.org/10.1186/s13059-021-02506-9

[36] Samanta, P., Cooke, S.F., McNulty, R., Hormoz, S., Rosenthal, A.: Probac-seq, a bacterial single-cell rna sequencing methodology using droplet microfluidics and large oligonucleotide probe sets. Nature Protocols, 1–28 (2024)

[37] Wolf, F.A., Angerer, P., Theis, F.J.: SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. **19**(1), 15 (2018) https://doi.org/10.1186/s13059-017-1382-0

[38] Virshup, I., Bredikhin, D., Heumos, L., Palla, G., Sturm, G., Gayoso, A., Kats, I., Koutrouli, M., Scverse Community, Berger, B., Pe'er, D., Regev, A., Teichmann, S.A., Finotello, F., Wolf, F.A., Yosef, N., Stegle, O., Theis, F.J.: The scverse project provides a computational ecosystem for single-cell omics data analysis. Nat. Biotechnol. **41**(5), 604–606 (2023) https://doi.org/10.1038/s41587-023-01733-8

[39] Vallejos, C.A., Risso, D., Scialdone, A., Dudoit, S., Marioni, J.C.: Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat. Methods **14**(6), 565–571 (2017) https://doi.org/10.1038/nmeth.4292

[40] Sina Booeshaghi, A., Hallgrímsdóttir, I.B., Gálvez-Merchán, Á., Pachter, L.: Depth normalization for single-cell genomics count data (2022). https://doi.org/10.1101/2022.05.06.490859 . https://www.biorxiv.org/content/10.1101/2022.05.06.490859v1

[41] Hafemeister, C., Satija, R.: Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. **20**(1), 296 (2019) https://doi.org/10.1186/s13059-019-1874-1

[42] McInnes, L., Healy, J., Saul, N., Großberger, L.: UMAP: Uniform manifold approximation and projection. J. Open Source Softw. **3**(29), 861 (2018) https://doi.org/10.21105/joss.00861

[43] Blondel, V., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. **2008** (2008) https://doi.org/10.1088/1742-5468/2008/10/P10008

[44] Traag, V.A., Waltman, L., Eck, N.J.: From louvain to leiden: guaranteeing well-connected communities. Sci. Rep. **9** (2019)

[45] Neufeld, A., Gao, L.L., Popp, J., Battle, A., Witten, D.: Inference after latent variable estimation for single-cell RNA sequencing data (2022) arXiv:2207.00554 [stat.ME]

[46] Zhang, J.M., Kamath, G.M., Tse, D.N.: Valid post-clustering differential analysis for Single-Cell RNA-Seq. Cell Systems **9**(4), 383–3926 (2019) https://doi.org/10.1016/j.cels.2019.07.012

[47] Sun, T., Song, D., Li, W.V., Li, J.J.: scdesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. Genome Biol. **22**(1), 163 (2021) https://doi.org/10.1186/s13059-021-02367-2

[48] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M. 3rd, Hao, Y., Stoeckius, M., Smibert, P., Satija, R.: Comprehensive integration of Single-Cell data. Cell **177**(7), 1888–190221 (2019) https://doi.org/10.1016/j.cell.2019.05.031

[49] Dubnau, D.: Genetic competence in bacillus subtilis. Microbiol. Rev. **55**(3), 395–424 (1991) https://doi.org/10.1128/mr.55.3.395-424.1991

[50] Ge, P., Scholl, D., Prokhorov, N.S., Avaylon, J., Shneider, M.M., Browning, C., Buth, S.A., Plattner, M., Chakraborty, U., Ding, K., Leiman, P.G., Miller, J.F., Zhou, Z.H.: Action of a minimal contractile bactericidal nanomachine. Nature **580**(7805), 658–662 (2020) https://doi.org/10.1038/s41586-020-2186-z

[51] Nakayama, K., Takashima, K., Ishihara, H., Shinomiya, T., Kageyama, M., Kanaya, S., Ohnishi, M., Murata, T., Mori, H., Hayashi, T.: The r-type pyocin of pseudomonas aeruginosa is related to P2 phage, and the f-type is related to lambda phage. Mol. Microbiol. **38**(2), 213–231 (2000) https://doi.org/10.1046/j.1365-2958.2000.02135.x

[52] Büttner, M., Ostner, J., Müller, C.L., Theis, F.J., Schubert, B.: scCODA is a bayesian model for compositional single-cell data analysis. Nat. Commun. **12**(1), 6876 (2021) https://doi.org/10.1038/s41467-021-27150-6

[53] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. **57**(1), 289–300 (1995)

[54] Fernandez, L., Rosvall, M., Normark, J., Fällman, M., Avican, K.: Co-PATHOgenex web application for assessing complex stress responses in pathogenic bacteria. Microbiol Spectr **12**(1), 0278123 (2024) https://doi.org/10.1128/spectrum.02781-23

[55] Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **15**(12), 550 (2014) https://doi.org/10.1186/s13059-014-0550-8

[56] Song, D., Wang, Q., Yan, G., Liu, T., Sun, T., Li, J.J.: scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics. Nat. Biotechnol. (2023) https://doi.org/10.1038/s41587-023-01772-1

[57] Ledoit, O., Wolf, M.: Honey, I Shrunk the Sample Covariance Matrix (2003). https://doi.org/10.2139/ssrn.433840 . https://papers.ssrn.com/abstract=433840

[58] Schäfer, J., Strimmer, K.: A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat. Appl. Genet. Mol. Biol. **4**, 32 (2005) https://doi.org/10.2202/1544-6115.1175

[59] Zappia, L., Phipson, B., Oshlack, A.: Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLoS Comput. Biol. **14**(6), 1–14 (2018) https://doi.org/10.1371/journal.pcbi.1006245

[60] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., Mulbregt, P., SciPy 1.0 Contributors: SciPy 1.0: fundamental algorithms for scientific computing in python. Nat. Methods **17**(3), 261–272 (2020) https://doi.org/10.1038/s41592-019-0686-2

[61] Harris, C.R., Millman, K.J., Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M.H., Brett, M., Haldane, A., Del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array programming with NumPy. Nature **585**(7825), 357–362 (2020) https://doi.org/10.1038/s41586-020-2649-2

[62] Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., Garber, R.L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L.L., Coulter, S.N., Folger, K.R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G.K., Wu, Z., Paulsen, I.T., Reizer, J., Saier, M.H., Hancock, R.E., Lory, S., Olson, M.V.: Complete genome sequence of pseudomonas aeruginosa PAO1, an opportunistic pathogen. Nature **406**(6799), 959–964 (2000) https://doi.org/10.1038/35023079

[63] Aframian, N., Bendori, S.O., Hen, S., Guler, P., Stokar-Avihail, A., Manor, E., Msaeed, K., Lipsman, V., Grinberg, I., Mahagna, A., Eldar, A.: Dormant phages communicate to control exit from lysogeny (2021). https://doi.org/10.1101/2021.09.20.460909 . https://www.biorxiv.org/content/10.1101/2021.09.20.460909v1.full

[64] Park, J., Dies, M., Lin, Y., Hormoz, S., Smith-Unna, S.E., Quinodoz, S., Hernández-Jiménez, M.J., Garcia-Ojalvo, J., Locke, J.C.W., Elowitz, M.B.: Molecular time sharing through dynamic pulsing in single cells. Cell Syst **6**(2), 216–22915 (2018) https://doi.org/10.1016/j.cels.2018.01.011

[65] Locke, J.C.W., Young, J.W., Fontes, M., Hernández Jiménez, M.J., Elowitz, M.B.: Stochastic pulse regulation in bacterial stress response. Science **334**(6054), 366–369 (2011) https://doi.org/10.1126/science.1208144

[66] Neidhardt, F.C., Bloch, P.L., Smith, D.F.: Culture medium for enterobacteria. Journal of bacteriology **119**(3), 736–747 (1974)

[67] Gummesson, B., Shah, S.A., Borum, A.S., Fessler, M., Mitarai, N., Sørensen, M.A., Svenningsen, S.L.: Valine-induced isoleucine starvation in escherichia coli k-12 studied by spike-in normalized rna sequencing. Frontiers in genetics **11**, 496392 (2020)

[68] Fessler, M., Gummesson, B., Charbon, G., Svenningsen, S.L., Sørensen, M.A.: Short-term kinetics of rRNA degradation in escherichia coli upon starvation for carbon, amino acid or phosphate. Mol. Microbiol. **113**(5), 951–963 (2020) https://doi.org/10.1111/mmi.14462

[69] Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., Gregory, M.T., Shuga, J., Montesclaros, L., Underwood, J.G., Masquelier, D.A., Nishimura, S.Y., Schnall-Levin, M., Wyatt, P.W., Hindson, C.M., Bharadwaj, R., Wong, A., Ness, K.D., Beppu, L.W., Deeg, H.J., McFarland, C., Loeb, K.R., Valente, W.J., Ericson, N.G., Stevens, E.A., Radich, J.P., Mikkelsen, T.S., Hindson, B.J., Bielas, J.H.: Massively parallel digital transcriptional profiling of single cells. Nat. Commun. **8**, 14049 (2017) https://doi.org/10.1038/ncomms14049

[70] Smith, T., Heger, A., Sudbery, I.: UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. Genome Res. **27**(3), 491–499 (2017) https://doi.org/10.1101/gr.209601.116

[71] Lambiotte, R., Delvenne, J.-C., Barahona, M.: Laplacian dynamics and multiscale modular structure in networks (2008) arXiv:0812.1770 [physics.soc-ph]

[72] Liu, S., Thennavan, A., Garay, J.P., Marron, J.S., Perou, C.M.: MultiK: an automated tool to determine optimal cluster numbers in single-cell RNA sequencing data. Genome Biol. **22**(1), 232 (2021) https://doi.org/10.1186/s13059-021-02445-5

[73] Tyler, S.R., Lozano-Ojalvo, D., Guccione, E., Schadt, E.E.: Anti-correlated feature selection prevents false discovery of subpopulations in scRNAseq. Nat. Commun. **15**(1), 699 (2024) https://doi.org/10.1038/s41467-023-43406-9

[74] Traag, V.A., Van Dooren, P., Nesterov, Y.: Narrow scope for resolution-limit-free community detection. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. **84**(1 Pt 2), 016114 (2011) https://doi.org/10.1103/PhysRevE.84.016114

[75] Seabold, S., Perktold, J.: Statsmodels: Econometric and statistical modeling with python. In: Proceedings of the 9th Python in Science Conference. SciPy, ??? (2010). https://doi.org/10.25080/majora-92bf1922-011 . https://conference.scipy.org/proceedings/scipy2010/seabold.html

[76] Badri, M., Kurtz, Z.D., Bonneau, R., Müller, C.L.: Shrinkage improves estimation of microbial associations under different normalization methods. NAR Genom Bioinform **2**(4), 100 (2020) https://doi.org/10.1093/nargab/lqaa100

[77] Kiefer, J.: Sequential minimax search for a maximum. Proc. Am. Math. Soc. **4**(3), 502–506 (1953) https://doi.org/10.1090/S0002-9939-1953-0055639-3

Appendix A Additional dataset analysis

This section contains biological interpretation of selected datasets that were not discussed in the main text.

A.1 BacSC reveals effects of DNA damage in *B.subtilis*

One more impression on how external factors can change the composition of bacterial cell types is provided by the *Bsub_damage_PB* dataset by comparing this data to the same species grown in minimal media without DNA damage. First, the PCA plot of the DNA-damaged population did not exhibit the characteristic separation into three subpopulations as observed in the *Bsub_minmed_PB* dataset (Figure C10A). Instead, the UMAP embedding showed a much more homogeneous population structure C10B) with six different subclusters, and one separate cell type (cluster 6).

This cell type again contained competent cells, as indicated by an overexpression of *com* genes (FDR=0.1, Figure C10E, F, Table E10) although in a much lower concentration than in the experiment without DNA damage (0.9% vs. 9.4% of analyzed population). For cell types 1 and 2, BacSC found many genes to be up- or downregulated, respectively, at an FDR level of 0.1. Cell type 4 showed an overexpression of genes related to subtilosin A production (*albE, albF, albC, albA, albD*), while cell types 3 and 5 showed an overexpression of genes related to the SPbeta prophage (*yomS, yomP, yomR, ...*), and prophage PBSX (*xtmA, xtmB, xkdE, xkdC, etc.*), albeit only at FDR levels larger than 0.5.

A.2 BacSC discovers a new cell type in *K. pneumoniae*

The *Klebs_untreated_BD* data contains 48,511 cells after quality control and is thus the largest experiment of our analyzed datasets, but also one of the most sparse (99.1% zero entries, Table E1). The PCA plot generated by BacSC (Figure C12A) showed a separation of many cells that were later clustered as cell type 1 (Figure C12B). This cell type showed higher sequencing depth (Figure C12D) and a larger number of unique expressed genes per cell on average (Figure C12C).

Clustering revealed three distinct subpopulations (Figure C12B). Cell type 1 showed a distinct set of genes that were upregulated at an FDR of 0.05 (Figure C12E, G; Table E12). This cell type comprised 2,194 cells and was characterized by IS903B transposase genes (*RS22855*, Figure C12F). This MGE subpopulation was already described in the original publication, but separated more clearly from the rest of the population in the UMAP generated by BacSC (Figure C12B).

Cell type 0 made up the bulk of the cell population (44,236 cells) and was distinguished from the other cell types by no expression of IS903B transposase genes. The analysis with BacSC also found another cell type (Cluster 2), which was not described by [13]. Similar to the high-ribosomal cell type discovered in *P. aeruginosa*, this subpopulation was mostly characterized by a higher expression of ribosomal genes (*rplP, rplC, rpoC*).

Appendix B Supplementary figures

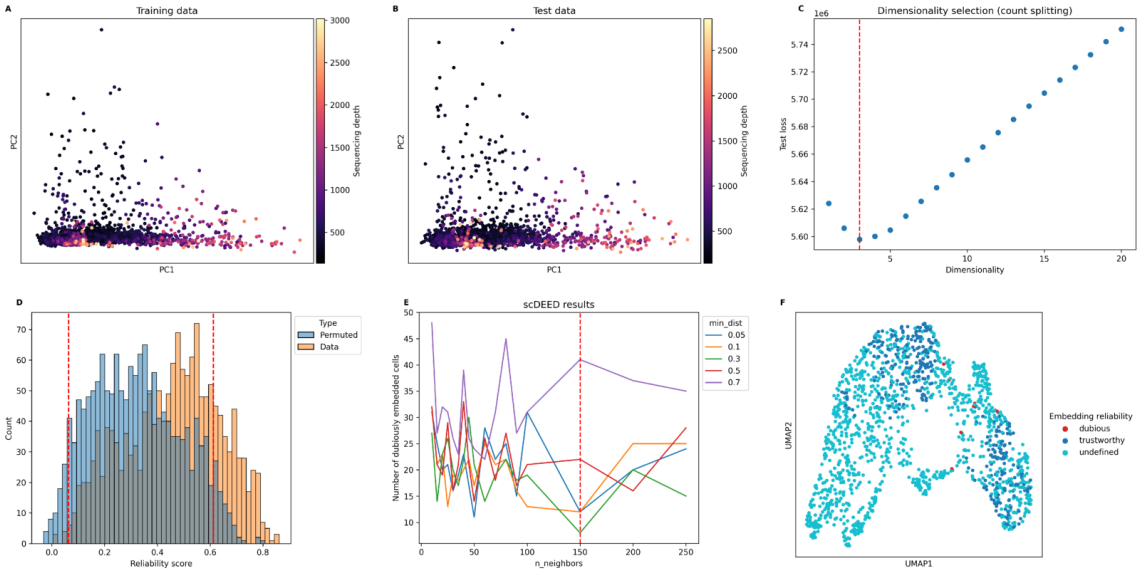


Fig. B1 Dimensionality reduction techniques in BacSC. All plots were generated for the *Pseudomonas_balanced_PB* dataset. Count splitting generates a training dataset **(A)** and test dataset **(B)** with similar PCA embeddings and count distribution. **(C)** The latent dimensionality k_{opt} of the dataset (dashed red line) is determined by minimizing the test loss. **(D)** Histogram of Null and target data reliability scores from scDEED. The dashed red lines denote the 5% and 95% quantiles of the distribution of null reliability scores ($n_{neighbors} = 150, min_{dist} = 0.3$). Cells with reliability scores smaller than the 5% quantile are marked as dubiously embedded, cells with reliability scores larger than the 95% quantile are marked as reliably embedded. **(E)** Number of dubiously embedded cells for each parameter combination tested in scDEED. The dashed red line indicates the chosen parameters $n_{neighbors} = 150, min_{dist} = 0.3$. **(F)** UMAP of the full dataset with parameters selected as in (E) and cells colored by their reliability classification from (D).

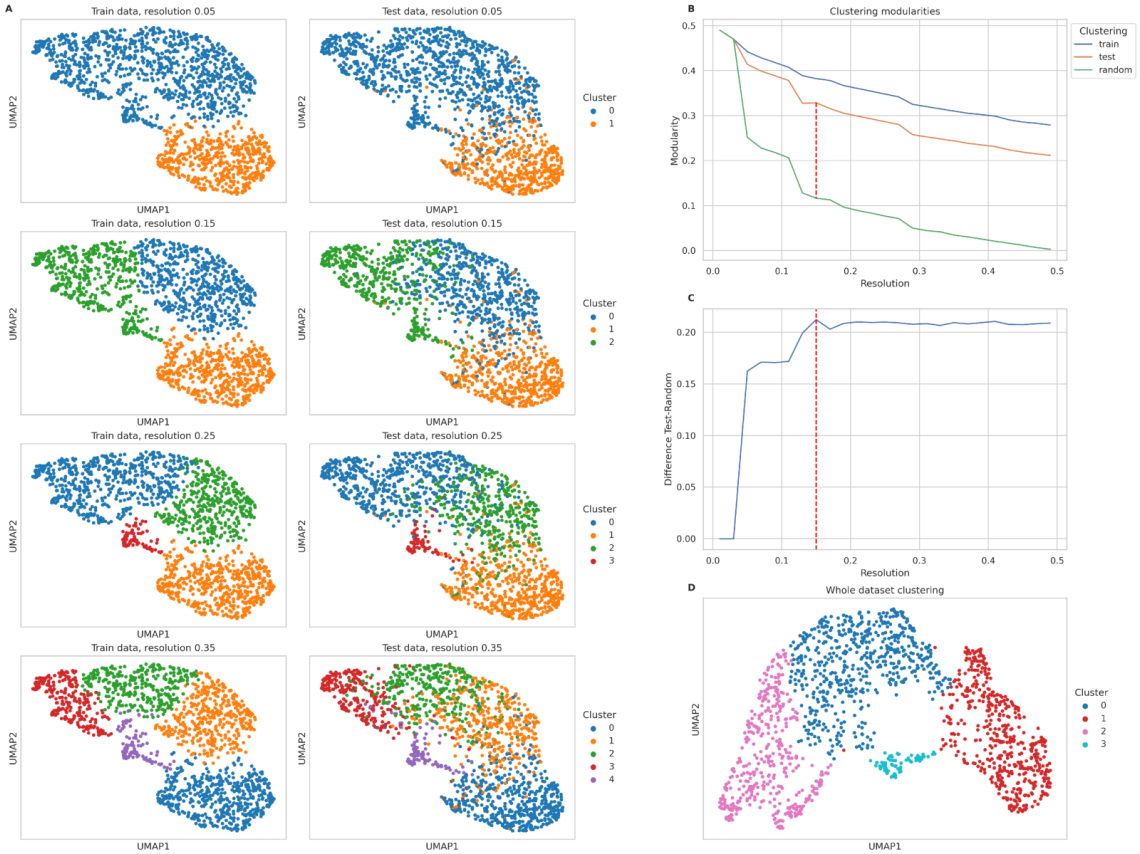


Fig. B2 Selection of clustering resolution in BacSC. All plots were generated for the *Pseudomonas_balanced_PB* dataset. **(A)** Clustering on train data (left column) for different resolutions, and applied to test data (right column). **(B)** Modularity scores of train data clustering on train data (blue), test data (orange), and of randomly shuffled clustering on test data (green) for all tested resolutions. The dashed line indicates the largest value of the gap statistic between test and random resolution. This resolution value is selected by BacSC. **(C)** Gap statistic between test and random resolution for all tested values of the resolution parameter. The dashed line indicates the chosen resolution res_{opt}. **(D)** UMAP of full dataset, clustered with the resolution parameter determined in (C) and (D).

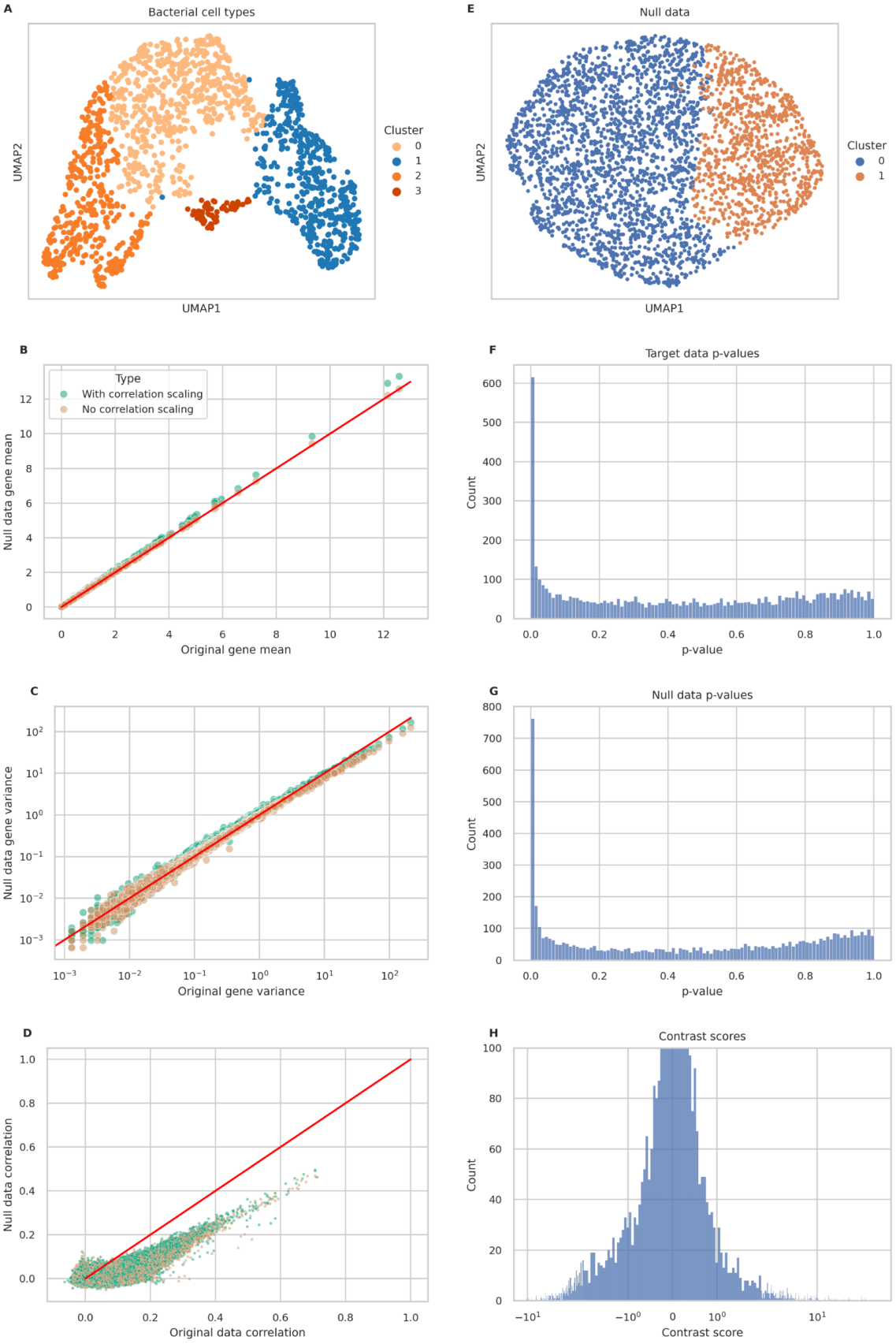


Fig. B3 Differential expression testing in BacSC. All plots were generated for the *Pseudomonas_balanced_PB* dataset. **(A)** UMAP of the target data. In this figure, the testing of cluster 1 (blue) against the union of all other clusters (orange) is shown. **(B)** Comparison of gene means for synthetic null data with and without correlation scaling to the original data. The red line indicates a perfect match. **(C)** Comparison of gene variances for synthetic null data with and without correlation scaling to the original data. The red line indicates a perfect match. **(D)** Comparison of empirical gene-gene correlations (shrunk by the procedure outlined in [76]) for synthetic null data with and without correlation scaling to the original data. Only a random subset of 100,000 correlations is shown for each type of synthetic data. The red line indicates a perfect match. **(E)** UMAP of processed null dataset with clustering into two subsets. **(F)** Histogram of p-values for testing cell type 1 against all other cell types on the original (target) data. **(G)** Histogram of p-values for testing cell type 1 against cell type 0 on the synthetic null data. **(H)** Histogram of contrast scores for testing cell type 1 against cell type 0. The y-axis was truncated at 100.

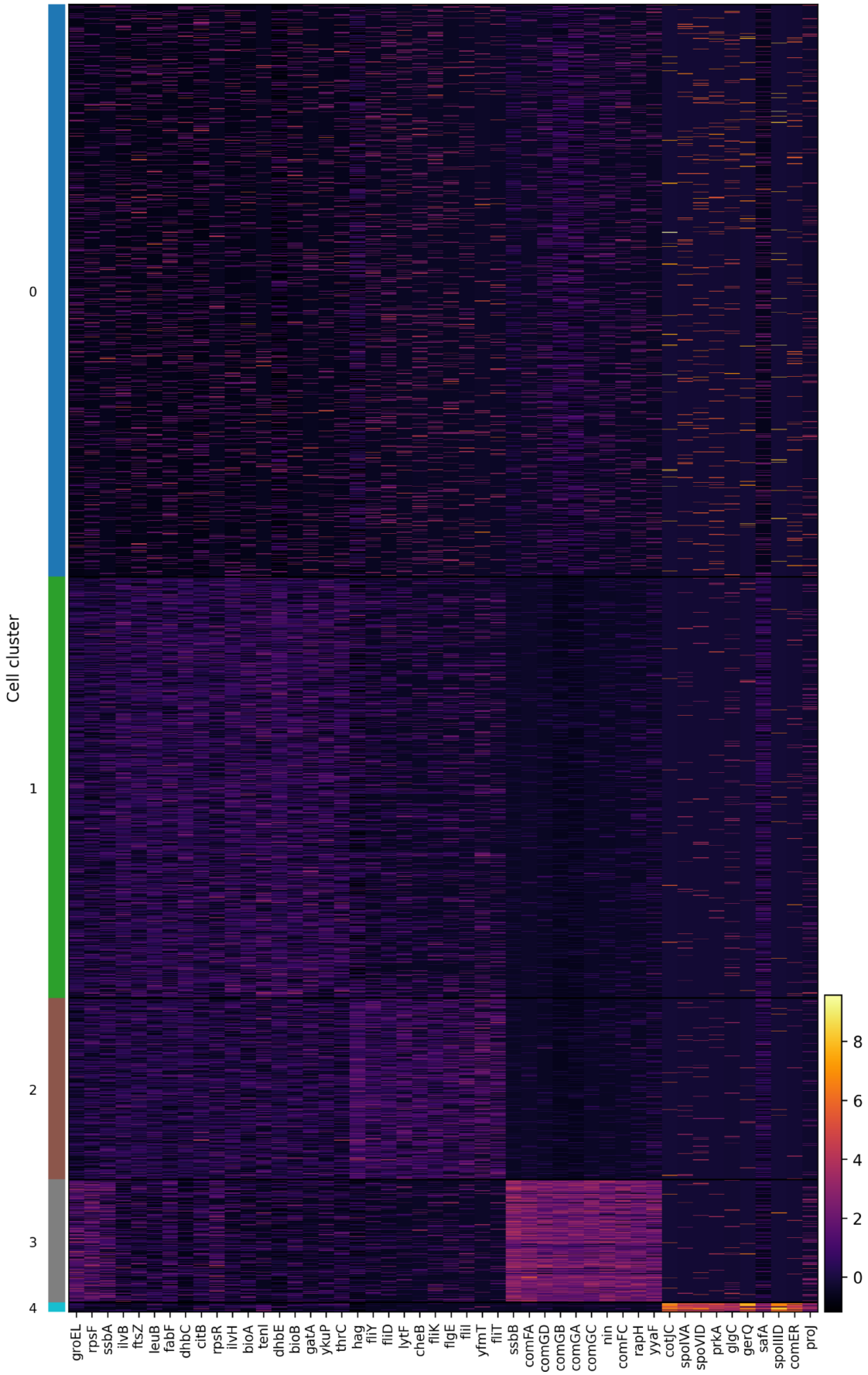


Fig. B4 Full heatmap of normalized gene expression for the *Bsub_minmed_PB* dataset. This figure extends Figure 3E by showing all cells. For each cluster, the 10 genes with the highest contrast scores are shown.

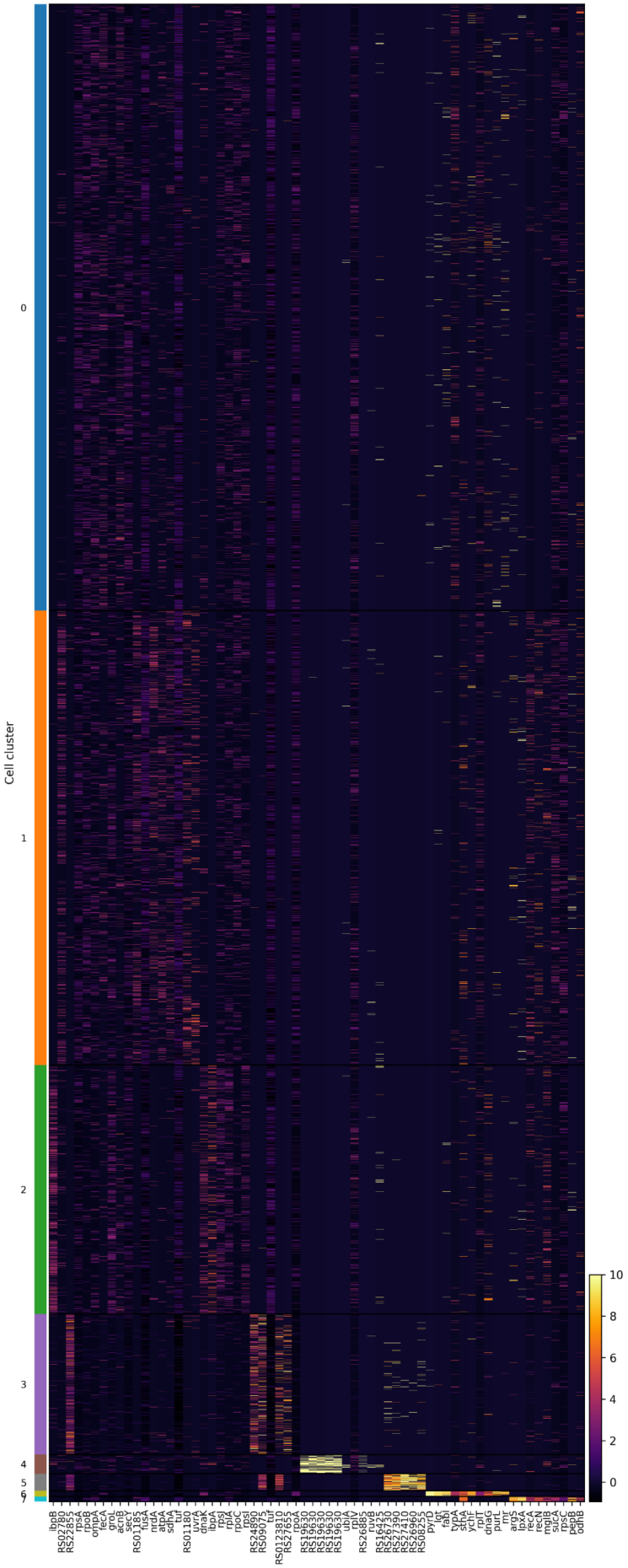


Fig. B5 Full heatmap of normalized gene expression for the *Klebs_antibiotics_PB* dataset. This figure extends Figure 4E by showing all cells. For each cluster, the 10 genes with the highest contrast scores are shown.

Appendix C Additional dataset analysis

This section contains results from BacSC in the style of figures 3 and 4 for all datasets shown in table 1 that were not already shown the main text.

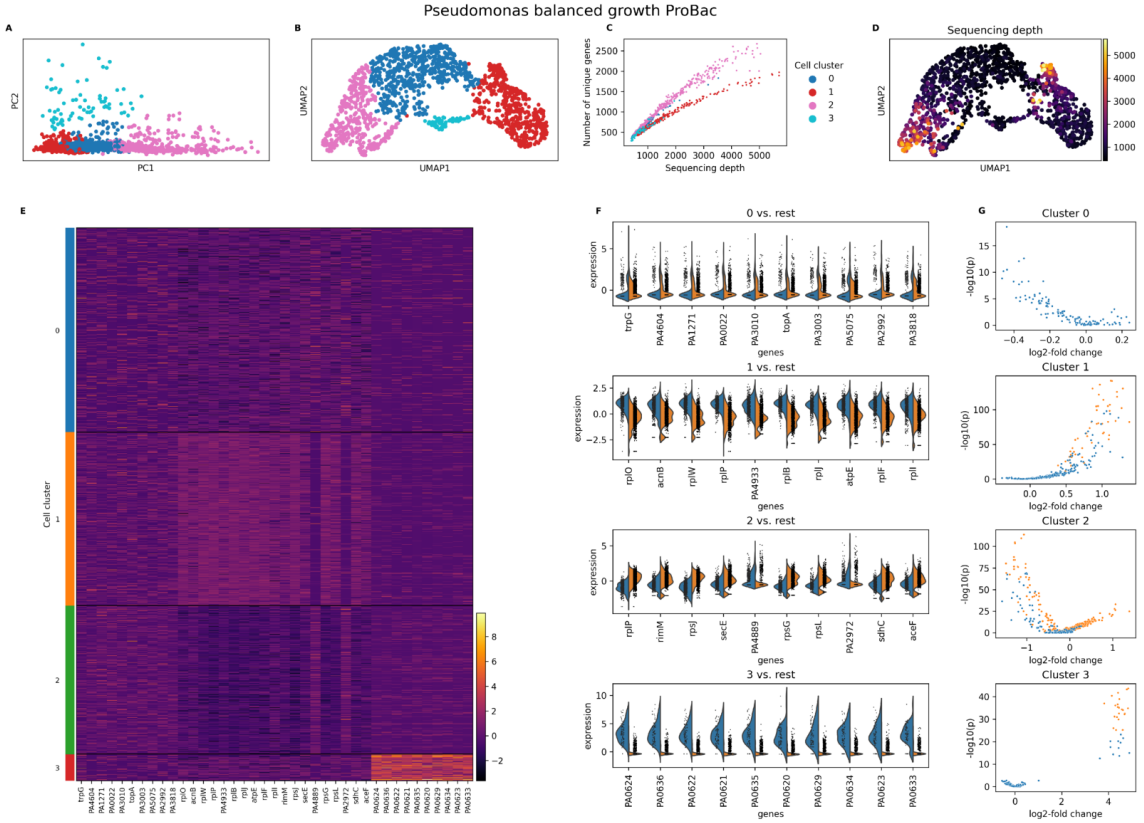


Fig. C6 Analysis of the *Pseudomonas_balanced_PB* dataset with BacSC. (A) Scatterplot of first two dimensions of PCA embedding with cell type clusters highlighted. **(B)** UMAP plot based on the parameters determined by BacSC, colored by cell cluster. **(C)** Scatterplot of sequencing depth versus number of unique genes per cell, colored by cell cluster. **(D)** Umap plot as in (B), colored by sequencing depth. **(E)** Heatmap of normalized gene expression for all cells and characteristic genes for cell types. For each cluster, the 10 genes with the highest contrast scores are shown. **(F)** Violin plots of normalized gene expression for DE tests of each cell type (blue) against the rest of the cell population (orange). For each cluster, the 10 genes with the highest contrast score are shown. Genes in (E) and (F) are annotated with gene symbols wherever possible, otherwise locus tags are shown. **(G)** Volcano plots for DE tests as in (F). The x-axis shows the log-fold change for gene expression, the y-axis shows the $-\log_{10}$-transformed (uncorrected) p-value. Only genes that are expressed in at least 80% of cells in the respective cluster are shown. Orange genes are differentially expressed at the $\alpha = 0.05$-level.

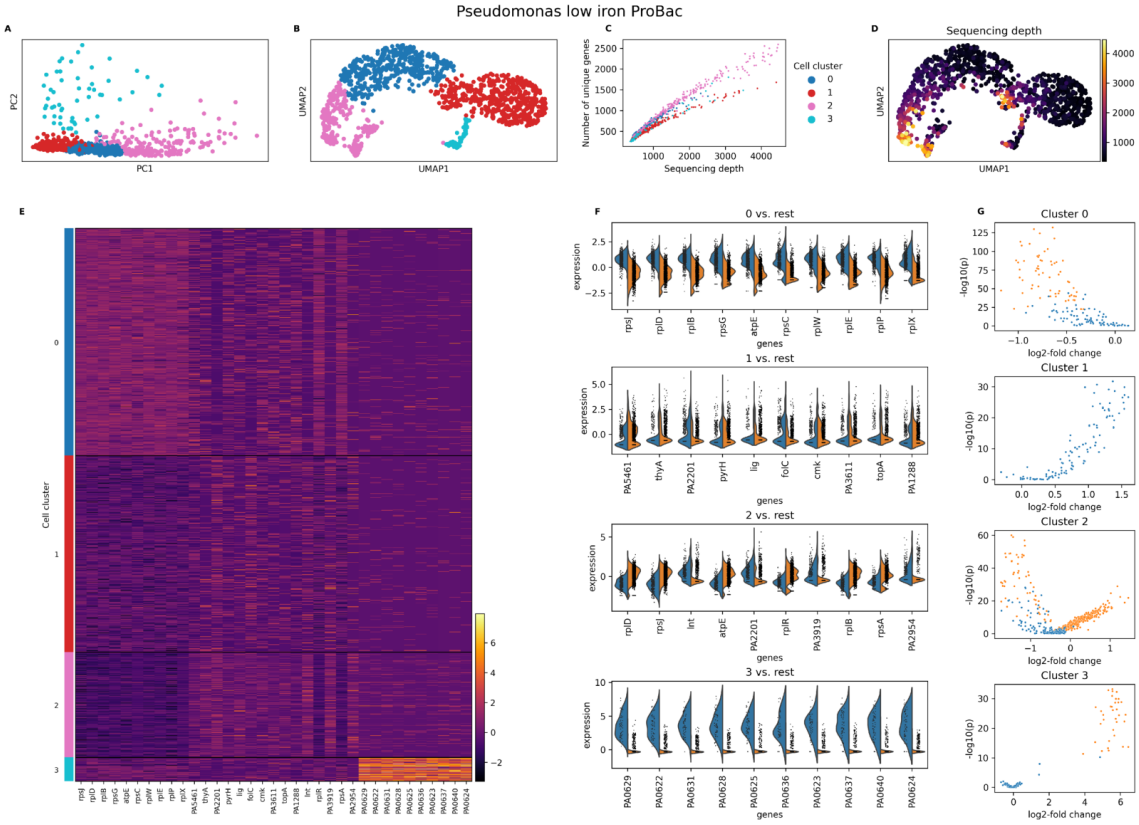


Fig. C7 Analysis of the *Pseudomonas_li_PB* dataset with BacSC. (A) Scatterplot of first two dimensions of PCA embedding with cell type clusters highlighted. **(B)** UMAP plot based on the parameters determined by BacSC, colored by cell cluster. **(C)** Scatterplot of sequencing depth versus number of unique genes per cell, colored by cell cluster. **(D)** Umap plot as in (B), colored by sequencing depth. **(E)** Heatmap of normalized gene expression for all cells and characteristic genes for cell types. For each cluster, the 10 genes with the highest contrast scores are shown. **(F)** Violin plots of normalized gene expression for DE tests of each cell type (blue) against the rest of the cell population (orange). For each cluster, the 10 genes with the highest contrast score are shown. Genes in (E) and (F) are annotated with gene symbols wherever possible, otherwise locus tags are shown. **(G)** Volcano plots for DE tests as in (F). The x-axis shows the log-fold change for gene expression, the y-axis shows the $-\log_{10}$-transformed (uncorrected) p-value. Only genes that are expressed in at least 80% of cells in the respective cluster are shown. Orange genes are differentially expressed at the $\alpha = 0.05$-level.

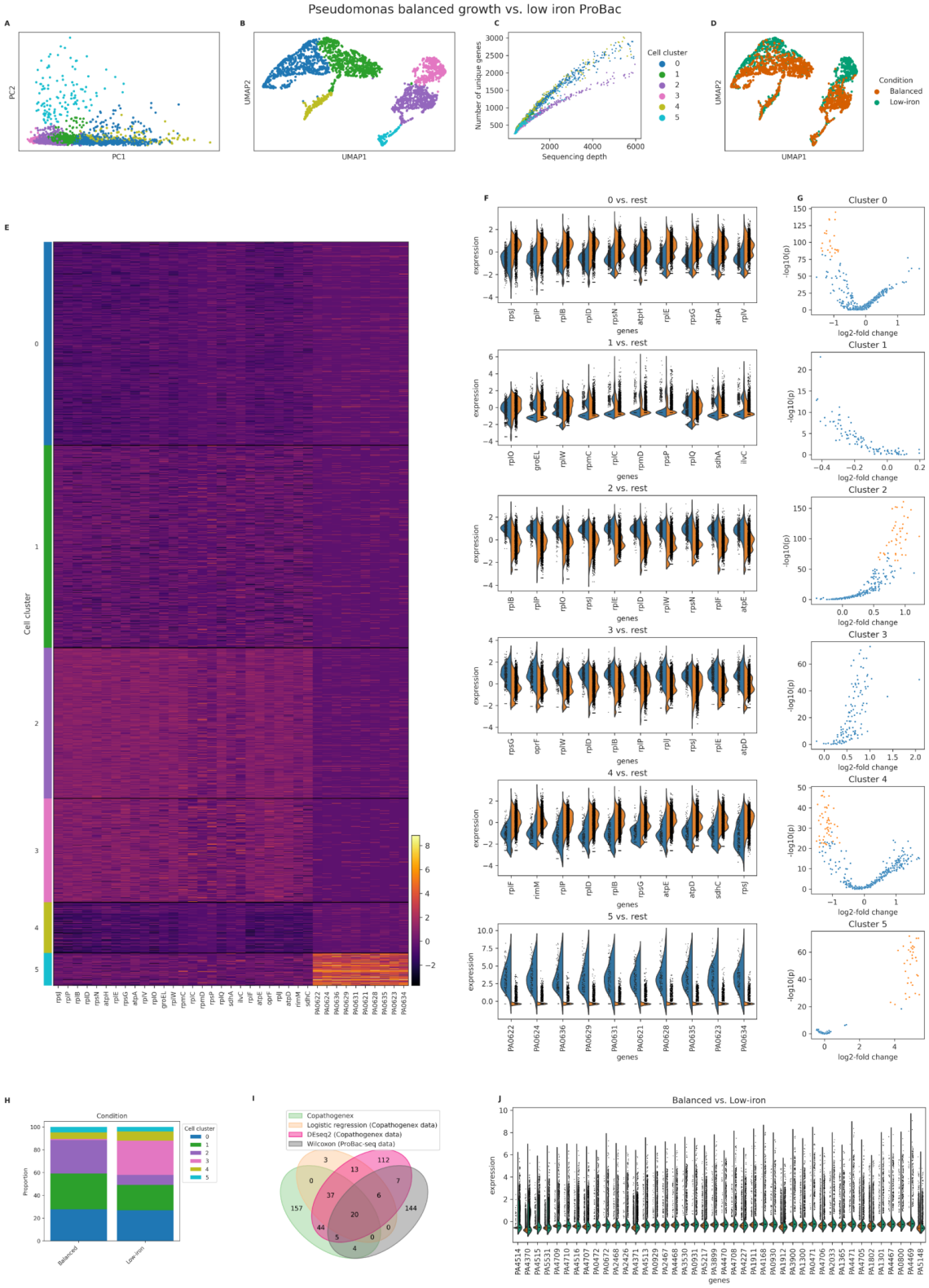


Fig. C8 Analysis of the combined *Pseudomonas_balanced_PB* and *Pseudomonas_li_PB* dataset with BacSC. **(A)** Scatterplot of first two dimensions of PCA embedding with cell type clusters highlighted. **(B)** UMAP plot based on the parameters determined by BacSC, colored by cell cluster. **(C)** Scatterplot of sequencing depth versus number of unique genes per cell, colored by cell cluster. **(D)** UMAP plot as in (B), colored by sample (growth condition). **(E)** Heatmap of normalized gene expression for all cells and characteristic genes for cell types. For each cluster, the 10 genes with the highest contrast scores are shown. **(F)** Violin plots of normalized gene expression for DE tests of each cell type (blue) against the rest of the cell population (orange). For each cluster, the 10 genes with the highest contrast score are shown. Genes in (E) and (F) are annotated with gene symbols wherever possible, otherwise locus tags are shown. **(G)** Volcano plots for DE tests as in (F). The x-axis shows the log-fold change for gene expression, the y-axis shows the $-\log_{10}$-transformed (uncorrected) p-value. Only genes that are expressed in at least 80% of cells in the respective cluster are shown. Orange genes are differentially expressed at the $\alpha = 0.05$-level. **(H)** Stacked barplot of cluster proportions for cells from each growth condition. **(I)** Venn diagram of differentially expressed genes found in Co-PATHOgenex and ProBac-seq data for Pseudomonas in balanced versus low-iron growth conditions. **(J)** Violin plots of differentially expressed genes in ProBac-seq and Co-PATHOgenex (at least one DE method, balanced vs. low-iron).

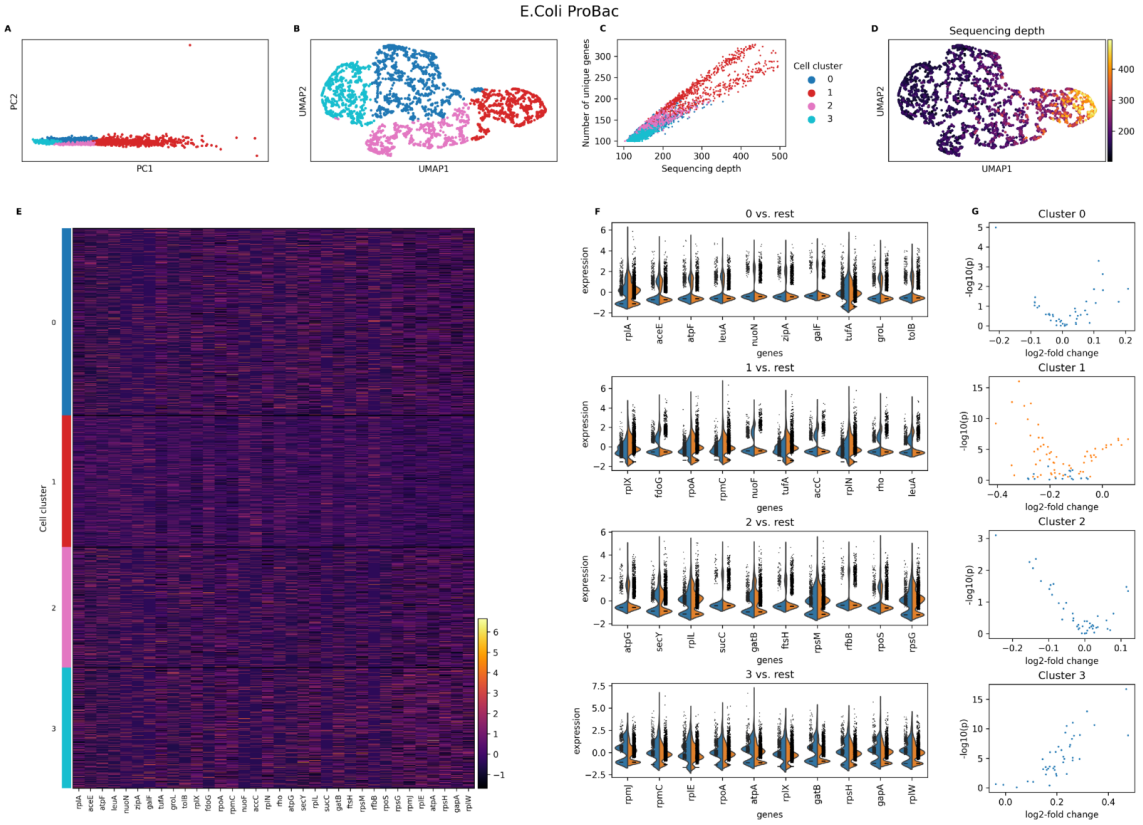


Fig. C9 Analysis of the *Ecoli_balanced_PB* dataset with BacSC. (A) Scatterplot of first two dimensions of PCA embedding with cell type clusters highlighted. **(B)** UMAP plot based on the parameters determined by BacSC, colored by cell cluster. **(C)** Scatterplot of sequencing depth versus number of unique genes per cell, colored by cell cluster. **(D)** Umap plot as in (B), colored by sequencing depth. **(E)** Heatmap of normalized gene expression for all cells and characteristic genes for cell types. For each cluster, the 10 genes with the highest contrast scores are shown. **(F)** Violin plots of normalized gene expression for DE tests of each cell type (blue) against the rest of the cell population (orange). For each cluster, the 10 genes with the highest contrast score are shown. **(G)** Volcano plots for DE tests as in (F). The x-axis shows the log-fold change for gene expression, the y-axis shows the $-\log_{10}$-transformed (uncorrected) p-value. Only genes that are expressed in at least 80% of cells in the respective cluster are shown. Orange genes are differentially expressed at the $\alpha = 0.05$-level.

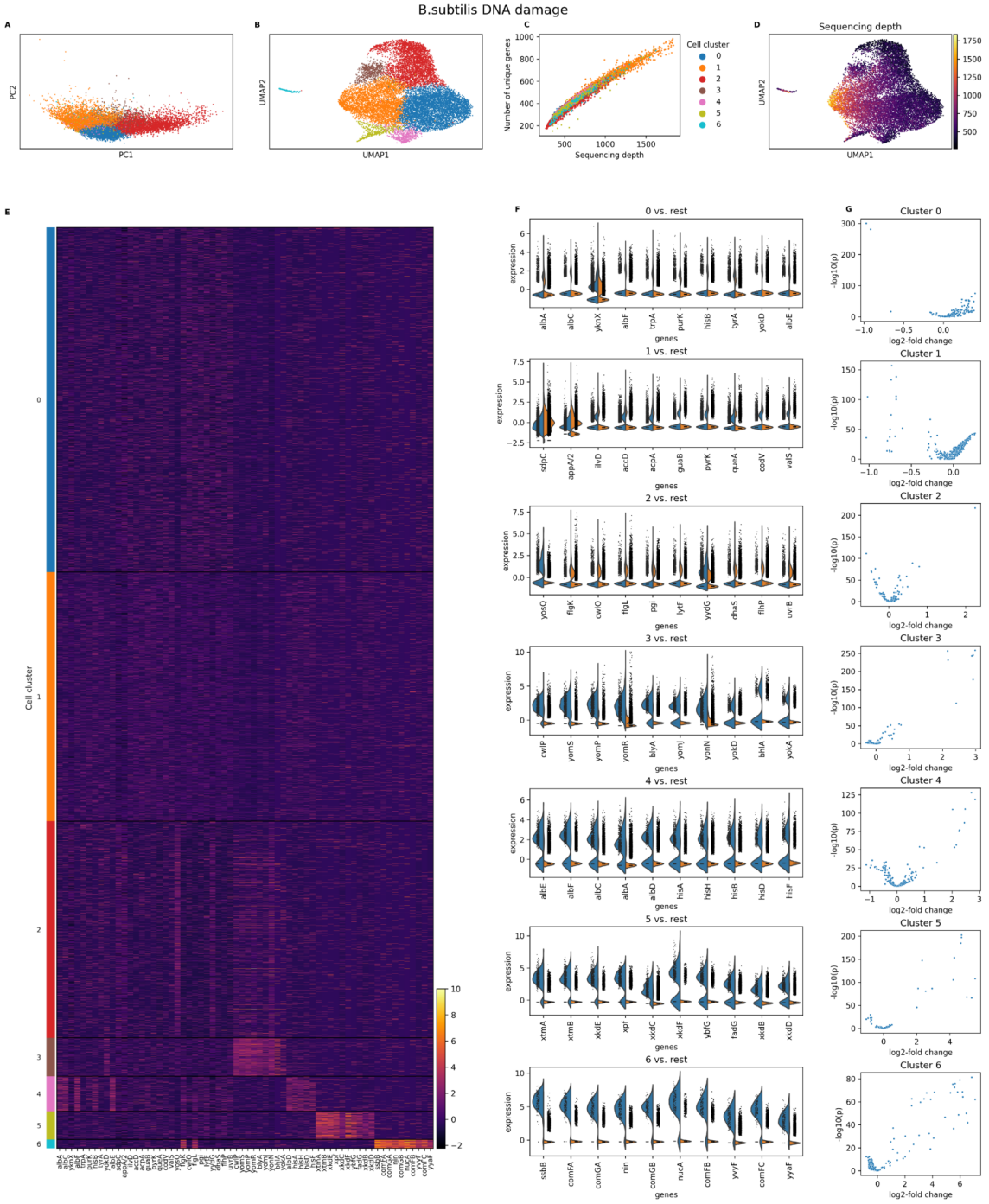


Fig. C10 Analysis of the *Bsub_damage_PB* dataset with BacSC. (A) Scatterplot of first two dimensions of PCA embedding with cell type clusters highlighted. **(B)** UMAP plot based on the parameters determined by BacSC, colored by cell cluster. **(C)** Scatterplot of sequencing depth versus number of unique genes per cell, colored by cell cluster. **(D)** Umap plot as in (B), colored by sequencing depth. **(E)** Heatmap of normalized gene expression for all cells and characteristic genes for cell types. For each cluster, the 10 genes with the highest contrast scores are shown. For each cluster, the 10 genes with the highest contrast scores are shown. **(F)** Violin plots of normalized gene expression for DE tests of each cell type (blue) against the rest of the cell population (orange). For each cluster, the 10 genes with the highest contrast score are shown. **(G)** Volcano plots for DE tests as in (F). The x-axis shows the log-fold change for gene expression, the y-axis shows the $-\log_{10}$-transformed (uncorrected) p-value. Only genes that are expressed in at least 80% of cells in the respective cluster are shown. Orange genes are differentially expressed at the $\alpha = 0.05$-level.

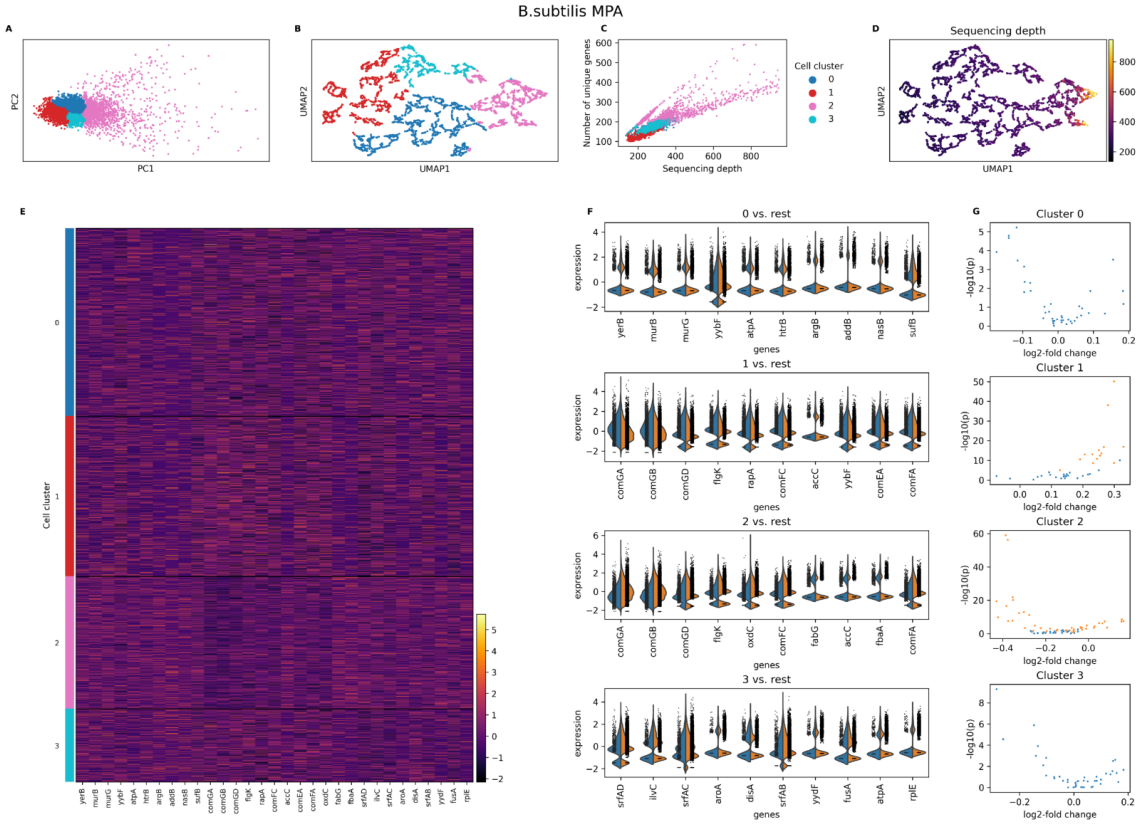


Fig. C11 Analysis of the *Bsub_MPA_PB* dataset with BacSC. (A) Scatterplot of first two dimensions of PCA embedding with cell type clusters highlighted. **(B)** UMAP plot based on the parameters determined by BacSC, colored by cell cluster. **(C)** Scatterplot of sequencing depth versus number of unique genes per cell, colored by cell cluster. **(D)** Umap plot as in (B), colored by sequencing depth. **(E)** Heatmap of normalized gene expression for all cells and characteristic genes for cell types. For each cluster, the 10 genes with the highest contrast scores are shown. **(F)** Violin plots of normalized gene expression for DE tests of each cell type (blue) against the rest of the cell population (orange). For each cluster, the 10 genes with the highest contrast score are shown. **(G)** Volcano plots for DE tests as in (F). The x-axis shows the log-fold change for gene expression, the y-axis shows the $-\log_{10}$-transformed (uncorrected) p-value. Only genes that are expressed in at least 80% of cells in the respective cluster are shown. Orange genes are differentially expressed at the $\alpha = 0.05$-level.

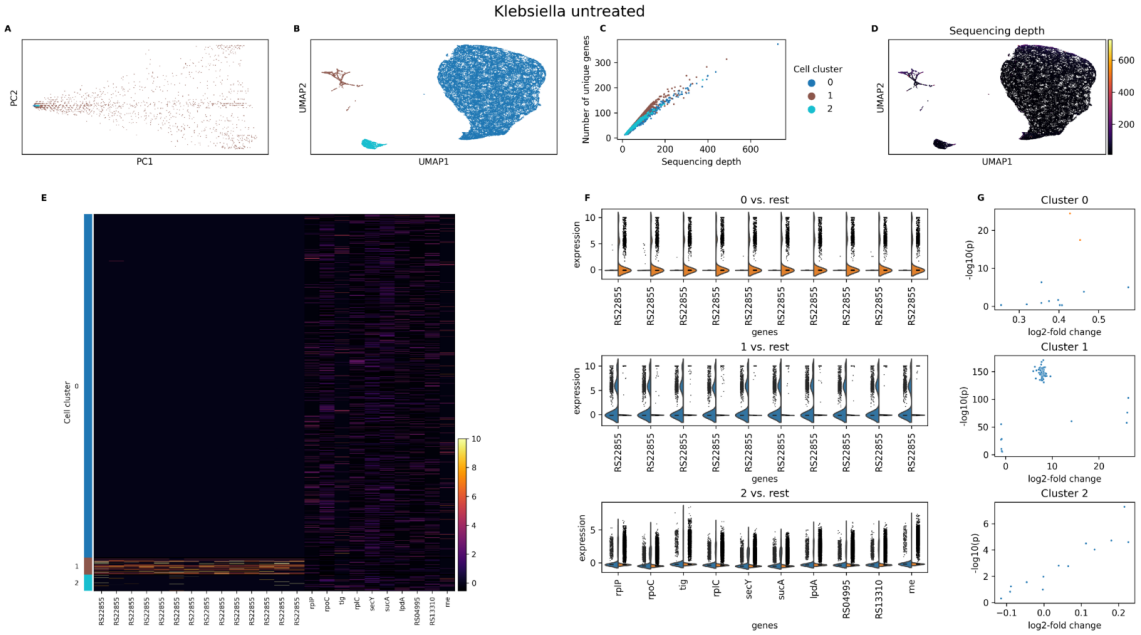


Fig. C12 Analysis of the *Klebs_untreated_BD* dataset with BacSC. (A) Scatterplot of first two dimensions of PCA embedding with cell type clusters highlighted. **(B)** UMAP plot based on the parameters determined by BacSC, colored by cell cluster. **(C)** Scatterplot of sequencing depth versus number of unique genes per cell, colored by cell cluster. **(D)** Umap plot as in (B), colored by sequencing depth. **(E)** Heatmap of normalized gene expression for all cells and characteristic genes for cell types. For each cluster, the 10 genes with the highest contrast scores are shown. **(F)** Violin plots of normalized gene expression for DE tests of each cell type (blue) against the rest of the cell population (orange). For each cluster, the 10 genes with the highest contrast score are shown. Genes in (E) and (F) are annotated with gene symbols wherever possible, otherwise locus tags are shown. **(G)** Volcano plots for DE tests as in (F). The x-axis shows the log-fold change for gene expression, the y-axis shows the $-\log_{10}$-transformed (uncorrected) p-value. Only genes that are expressed in at least 50% of cells in the respective cluster are shown. Orange genes are differentially expressed at the $\alpha = 0.05$-level.

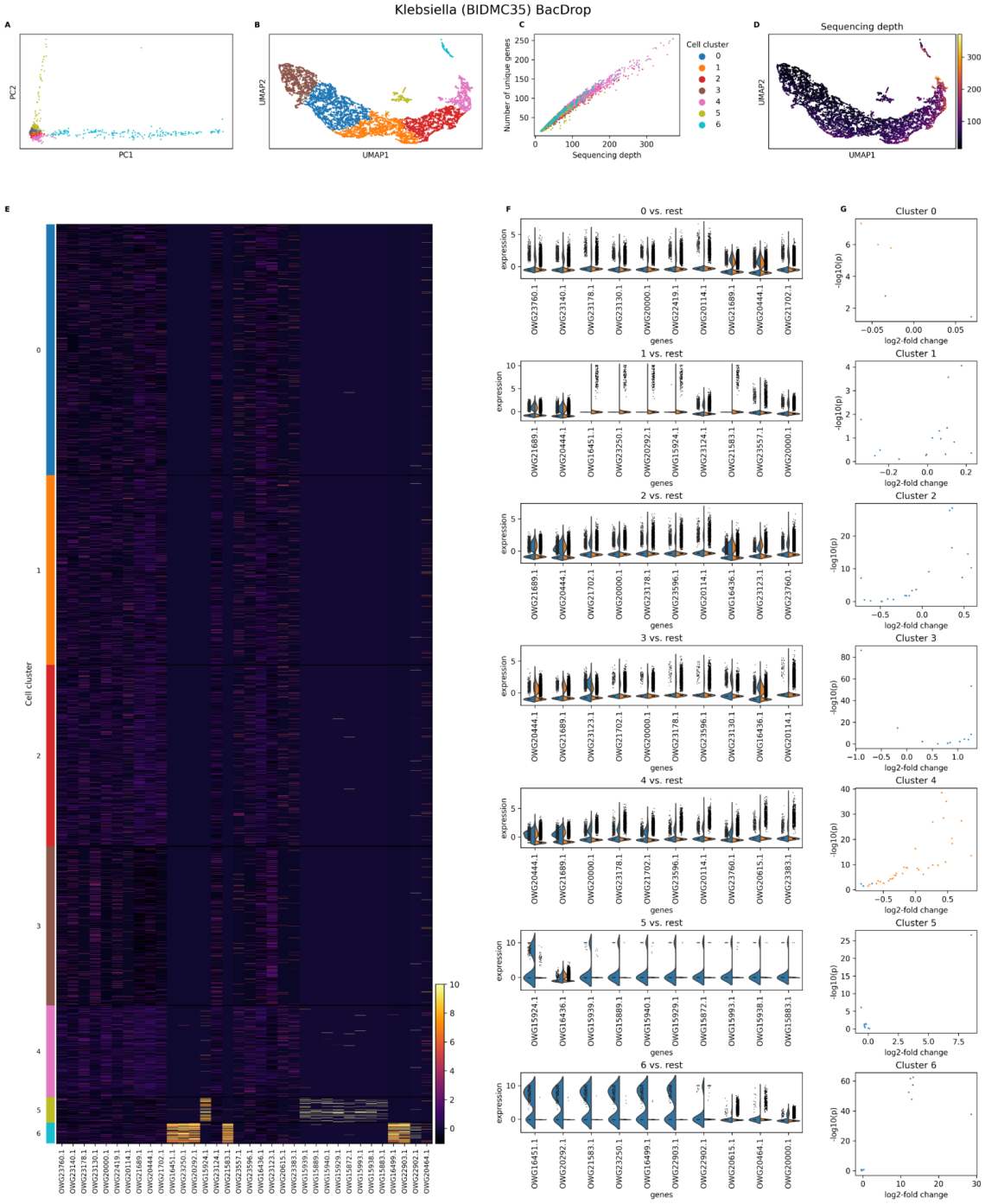


Fig. C13 Analysis of the *Klebs_BIDMC35_BD* dataset with BacSC. (A) Scatterplot of first two dimensions of PCA embedding with cell type clusters highlighted. **(B)** UMAP plot based on the parameters determined by BacSC, colored by cell cluster. **(C)** Scatterplot of sequencing depth versus number of unique genes per cell, colored by cell cluster. **(D)** Umap plot as in (B), colored by sequencing depth. **(E)** Heatmap of normalized gene expression for all cells and characteristic genes for cell types. For each cluster, the 10 genes with the highest contrast scores are shown. **(F)** Violin plots of normalized gene expression for DE tests of each cell type (blue) against the rest of the cell population (orange). For each cluster, the 10 genes with the highest contrast score are shown. Genes in (E) and (F) are annotated with gene symbols wherever possible, otherwise locus tags are shown. **(G)** Volcano plots for DE tests as in (F). The x-axis shows the log-fold change for gene expression, the y-axis shows the $-\log_{10}$-transformed (uncorrected) p-value. Only genes that are expressed in at least 50% of cells in the respective cluster are shown. Orange genes are differentially expressed at the $\alpha = 0.05$-level.

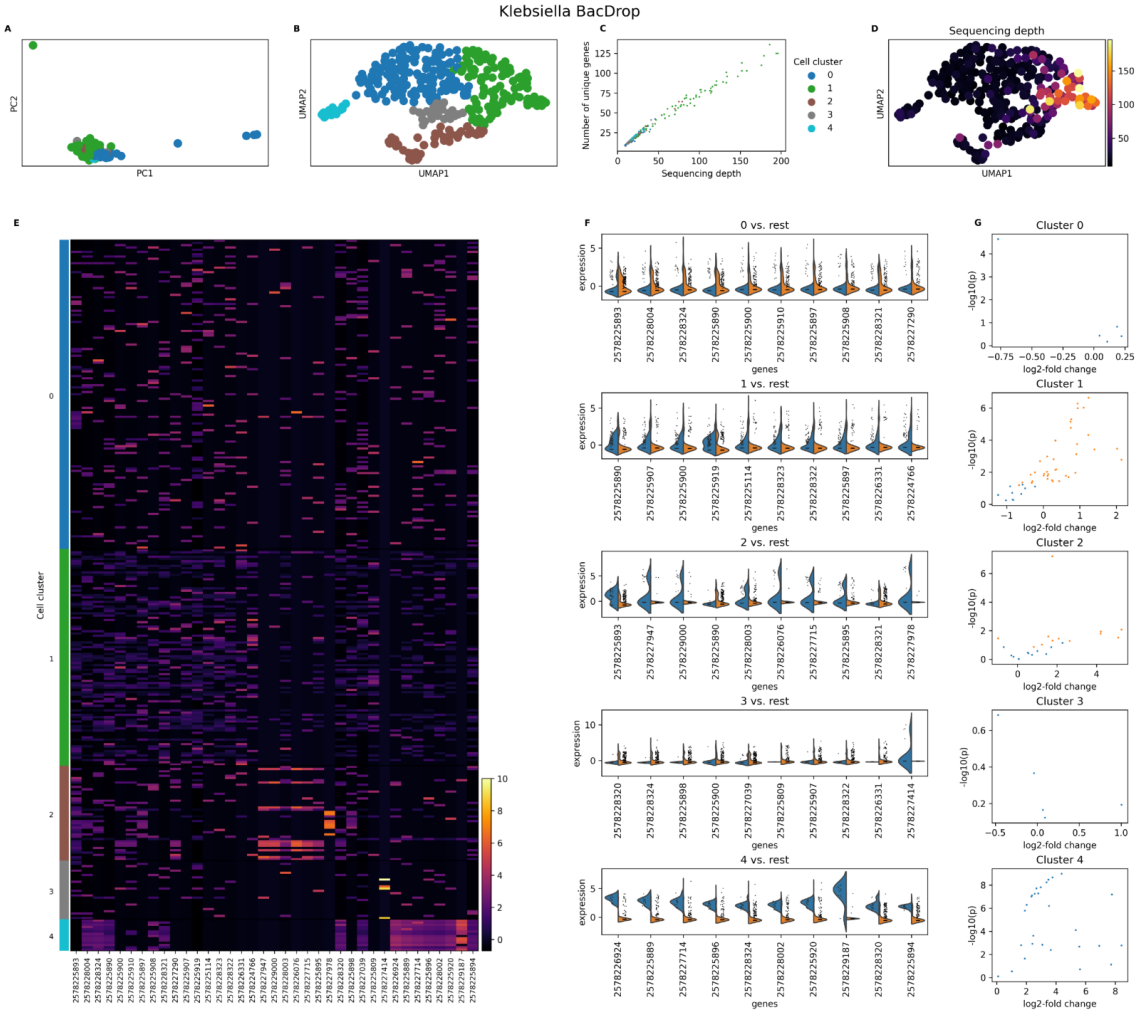


Fig. C14 Analysis of the *Klebs_4species_BD* dataset with BacSC. (A) Scatterplot of first two dimensions of PCA embedding with cell type clusters highlighted. **(B)** UMAP plot based on the parameters determined by BacSC, colored by cell cluster. **(C)** Scatterplot of sequencing depth versus number of unique genes per cell, colored by cell cluster. **(D)** Umap plot as in (B), colored by sequencing depth. **(E)** Heatmap of normalized gene expression for all cells and characteristic genes for cell types. For each cluster, the 10 genes with the highest contrast scores are shown. **(F)** Violin plots of normalized gene expression for DE tests of each cell type (blue) against the rest of the cell population (orange). For each cluster, the 10 genes with the highest contrast score are shown. Genes in (E) and (F) are annotated with gene symbols wherever possible, otherwise locus tags are shown. **(G)** Volcano plots for DE tests as in (F). The x-axis shows the log-fold change for gene expression, the y-axis shows the $-\log_{10}$-transformed (uncorrected) p-value. Only genes that are expressed in at least 50% of cells in the respective cluster are shown. Orange genes are differentially expressed at the $\alpha = 0.05$-level.

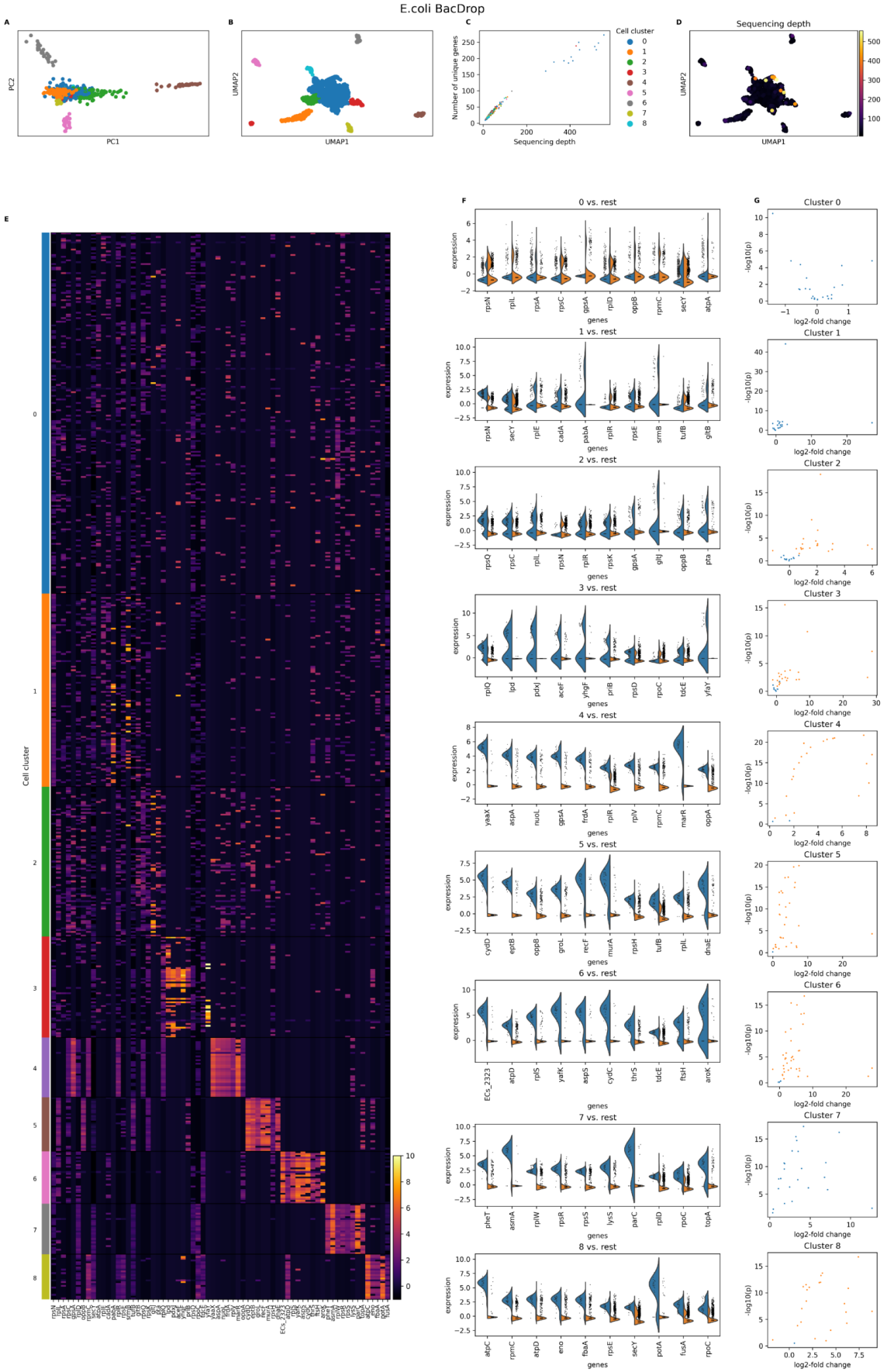


Fig. C15 Analysis of the *Ecoli_4species_BD* dataset with BacSC. (A) Scatterplot of first two dimensions of PCA embedding with cell type clusters highlighted. **(B)** UMAP plot based on the parameters determined by BacSC, colored by cell cluster. **(C)** Scatterplot of sequencing depth versus number of unique genes per cell, colored by cell cluster. **(D)** Umap plot as in (B), colored by sequencing depth. **(E)** Heatmap of normalized gene expression for all cells and characteristic genes for cell types. For each cluster, the 10 genes with the highest contrast scores are shown. **(F)** Violin plots of normalized gene expression for DE tests of each cell type (blue) against the rest of the cell population (orange). For each cluster, the 10 genes with the highest contrast score are shown. Genes in (E) and (F) are annotated with gene symbols wherever possible, otherwise locus tags are shown. **(G)** Volcano plots for DE tests as in (F). The x-axis shows the log-fold change for gene expression, the y-axis shows the $-\log_{10}$-transformed (uncorrected) p-value. Only genes that are expressed in at least 50% of cells in the respective cluster are shown. Orange genes are differentially expressed at the $\alpha = 0.05$-level.

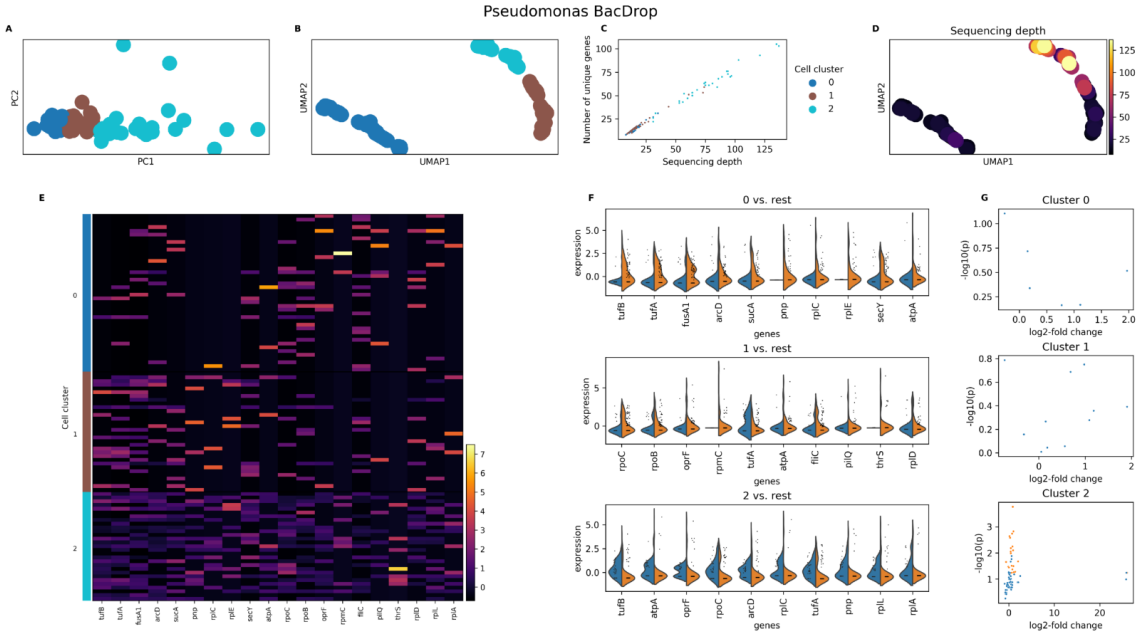


Fig. C16 Analysis of the *Pseudomonas_4species_BD* dataset with BacSC. (A) Scatterplot of first two dimensions of PCA embedding with cell type clusters highlighted. **(B)** UMAP plot based on the parameters determined by BacSC, colored by cell cluster. **(C)** Scatterplot of sequencing depth versus number of unique genes per cell, colored by cell cluster. **(D)** Umap plot as in (B), colored by sequencing depth. **(E)** Heatmap of normalized gene expression for all cells and characteristic genes for cell types. For each cluster, the 10 genes with the highest contrast scores are shown. **(F)** Violin plots of normalized gene expression for DE tests of each cell type (blue) against the rest of the cell population (orange). For each cluster, the 10 genes with the highest contrast score are shown. Genes in (E) and (F) are annotated with gene symbols wherever possible, otherwise locus tags are shown. **(G)** Volcano plots for DE tests as in (F). The x-axis shows the log-fold change for gene expression, the y-axis shows the $-\log_{10}$-transformed (uncorrected) p-value. Only genes that are expressed in at least 50% of cells in the respective cluster are shown. Orange genes are differentially expressed at the $\alpha = 0.05$-level.

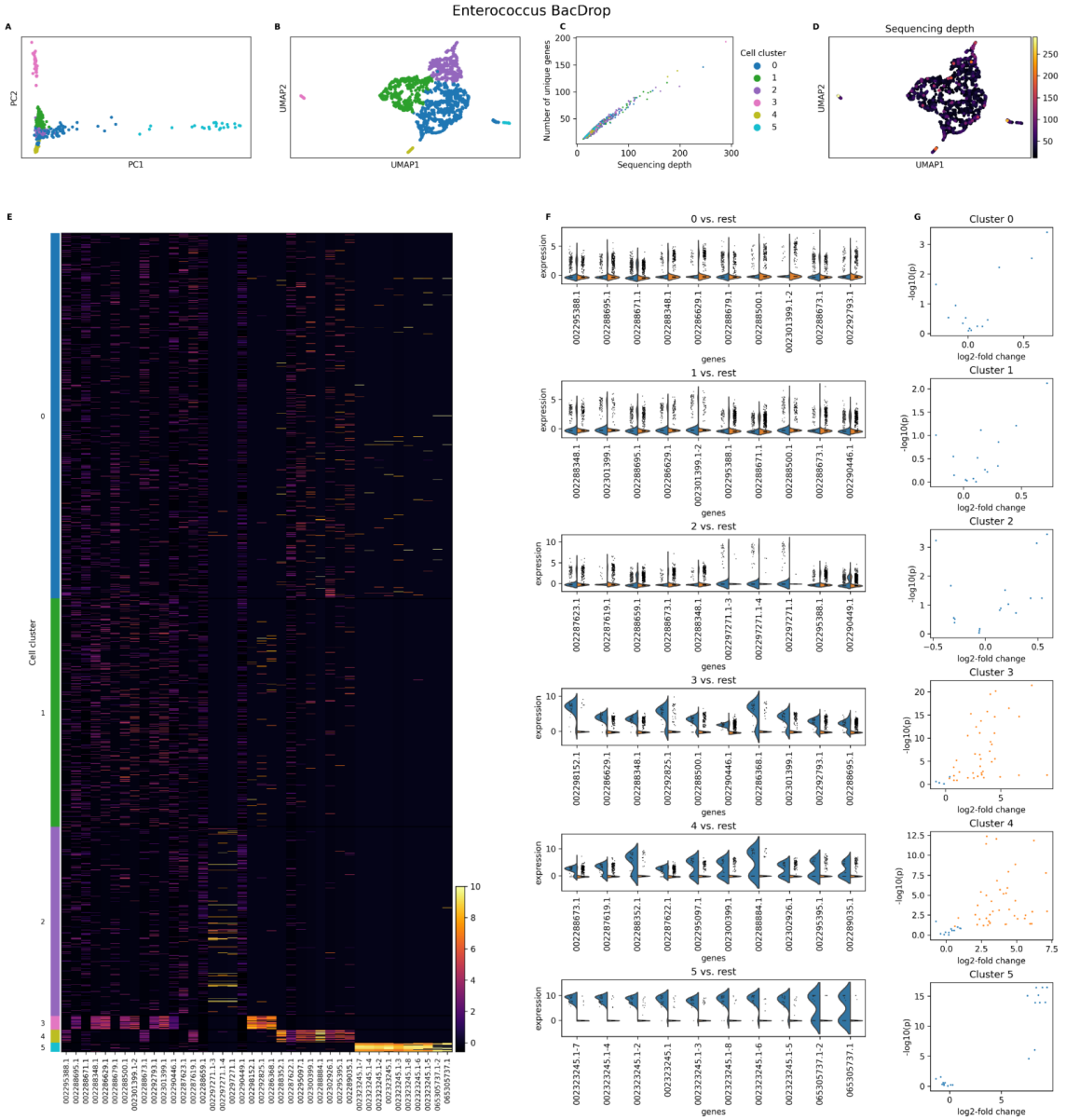


Fig. C17 Analysis of the *Efaecium_4species_BD* dataset with BacSC. (A) Scatterplot of first two dimensions of PCA embedding with cell type clusters highlighted. **(B)** UMAP plot based on the parameters determined by BacSC, colored by cell cluster. **(C)** Scatterplot of sequencing depth versus number of unique genes per cell, colored by cell cluster. **(D)** Umap plot as in (B), colored by sequencing depth. **(E)** Heatmap of normalized gene expression for all cells and characteristic genes for cell types. For each cluster, the 10 genes with the highest contrast scores are shown. **(F)** Violin plots of normalized gene expression for DE tests of each cell type (blue) against the rest of the cell population (orange). For each cluster, the 10 genes with the highest contrast score are shown. Genes in (E) and (F) are annotated with gene symbols wherever possible, otherwise locus tags are shown. **(G)** Volcano plots for DE tests as in (F). The x-axis shows the log-fold change for gene expression, the y-axis shows the $-\log_{10}$-transformed (uncorrected) p-value. Only genes that are expressed in at least 50% of cells in the respective cluster are shown. Orange genes are differentially expressed at the $\alpha = 0.05$-level.

Appendix D Diagnostic plots for all datasets

This section contains a selection of diagnostic plots from BacSC for each dataset from table 1.

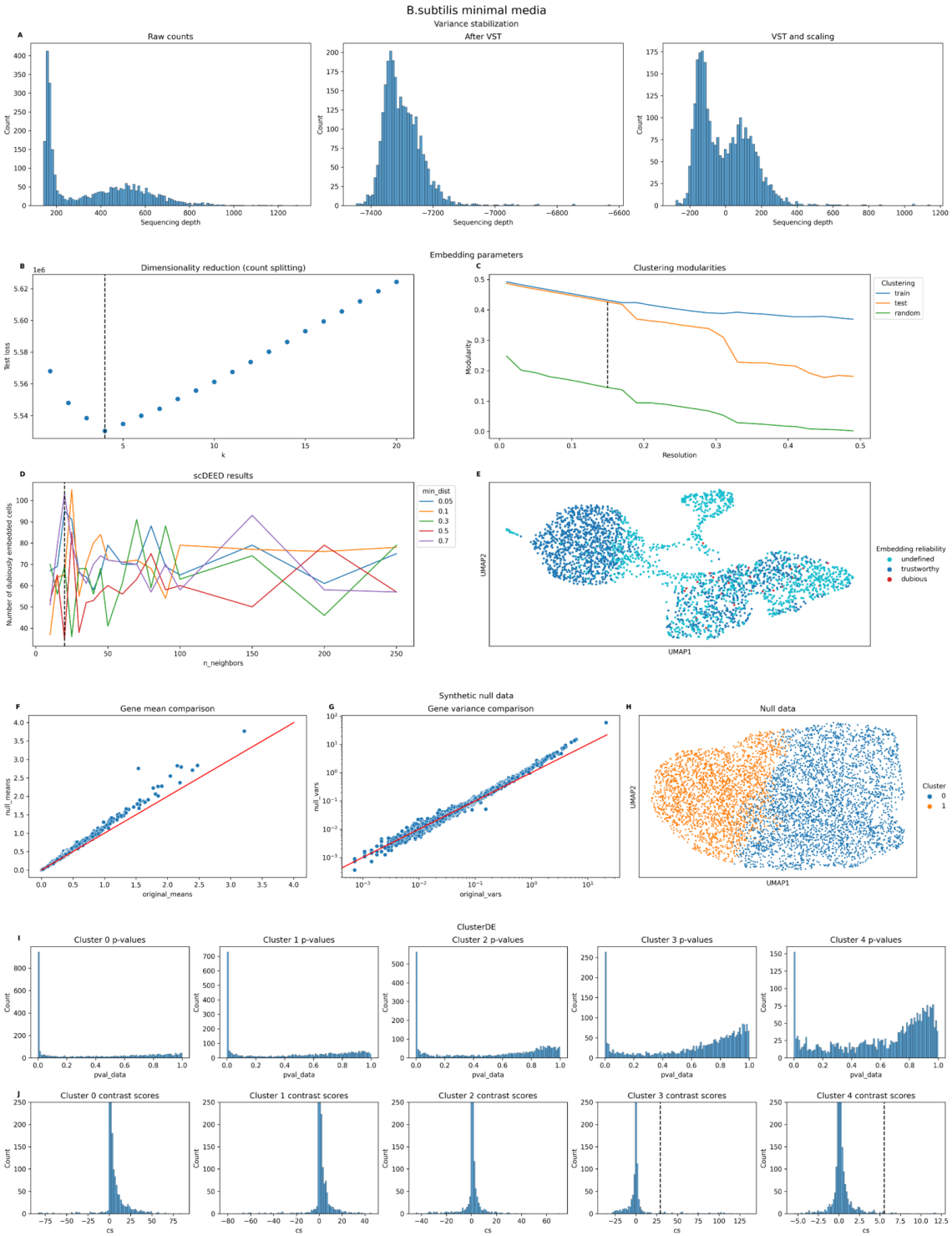


Fig. D18 Diagnostic plots generated during the analysis of the *Bsub_minmed_PB* dataset with BacSC. (A) Histograms of cell sequencing depth for raw gene expression data (left), after variance-stabilizing transform (middle), and after gene-wise scaling (right). **(B)** Selection of latent dimensionality through count splitting. The dashed line indicates k_{opt} selected by BacSC. **(C)** Selection of clustering resolution. The dashed line indicates the maximal gap in modularity between the training data Leiden clustering applied to the test data and a random clustering on the test data. **(D)** Selection of $n_{neighbors}$ and min_{dist} parameters for UMAP embedding with scDEED. The dashed black line shows the value of $n_{neighbors}$ selected by BacSC. **(E)** UMAP embedding generated with BacSC-selected parameters, colored by embedding reliability determined by scDEED with the same parameters. **(F)** Comparison of gene means of original and synthetic null data for DE testing. **(G)** Comparison of gene variances of original and synthetic null data for DE testing. **(H)** UMAP of synthetic null data for DE testing of each cell type against all other cells, colored by the two-group clustering determined for calculation of contrast scores. **(I)** Histograms of uncorrected p-values for DE testing of each cell type against all other cells. **(J)** Histograms of contrast scores for DE testing of each cell type against all other cells. The dashed lines indicate significance threshold values at the FDR level $\alpha = 0.05$.

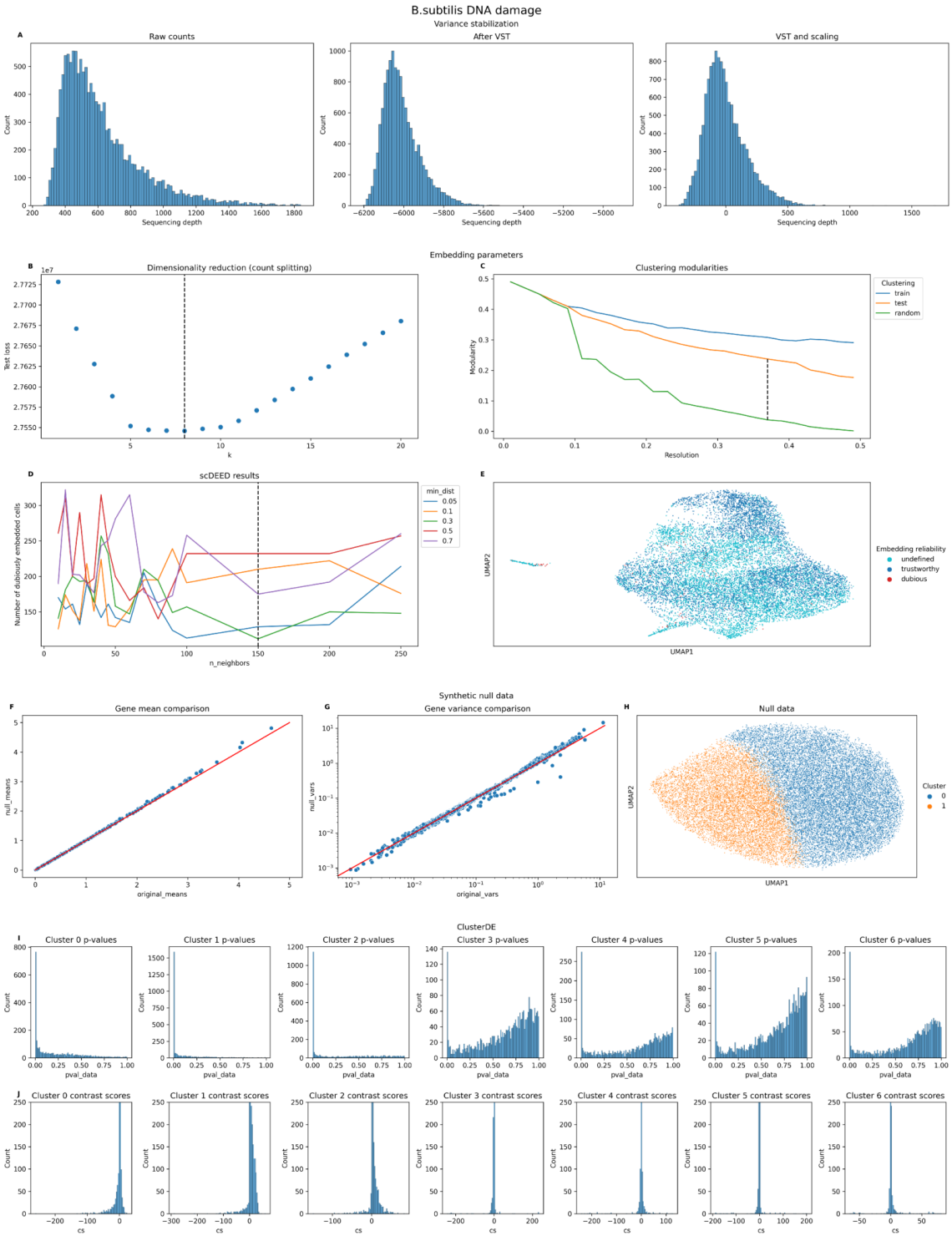


Fig. D19 Diagnostic plots generated during the analysis of the *Bsub_damage_PB* dataset with BacSC. (A) Histograms of cell sequencing depth for raw gene expression data (left), after variance-stabilizing transform (middle), and after gene-wise scaling (right). **(B)** Selection of latent dimensionality through count splitting. The dashed line indicates k_{opt} selected by BacSC. **(C)** Selection of clustering resolution. The dashed line indicates the maximal gap in modularity between the training data Leiden clustering applied to the test data and a random clustering on the test data. **(D)** Selection of $n_{neighbors}$ and min_{dist} parameters for UMAP embedding with scDEED. The dashed black line shows the value of $n_{neighbors}$ selected by BacSC. **(E)** UMAP embedding generated with BacSC-selected parameters, colored by embedding reliability determined by scDEED with the same parameters. **(F)** Comparison of gene means of original and synthetic null data for DE testing. **(G)** Comparison of gene variances of original and synthetic null data for DE testing. **(H)** UMAP of synthetic null data for DE testing of each cell type against all other cells, colored by the two-group clustering determined for calculation of contrast scores. **(I)** Histograms of uncorrected p-values for DE testing of each cell type against all other cells. **(J)** Histograms of contrast scores for DE testing of each cell type against all other cells. The dashed lines indicate significance threshold values at the FDR level $\alpha = 0.05$.

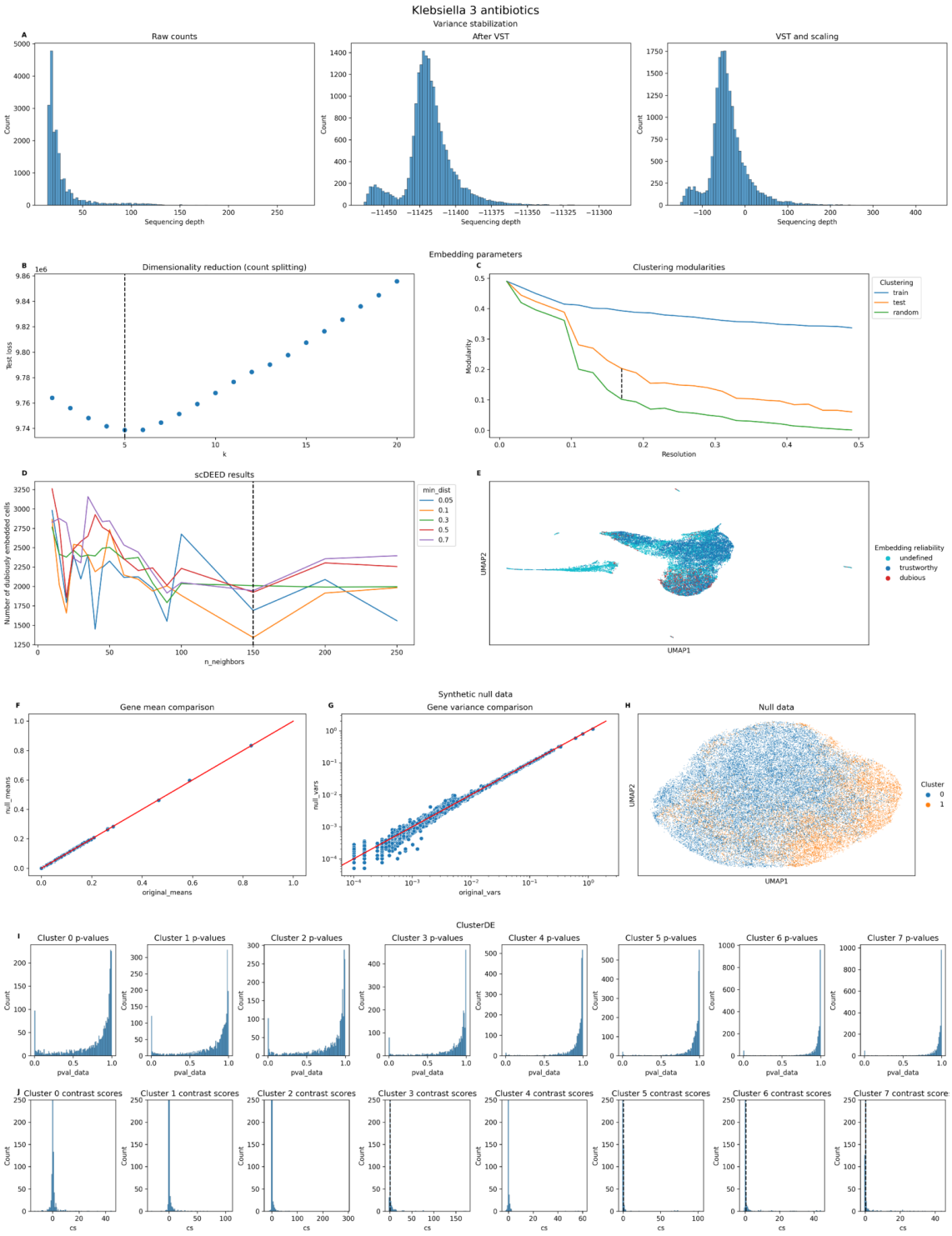


Fig. D20 Diagnostic plots generated during the analysis of the *Klebs_antibiotics_BD* dataset with BacSC. **(A)** Histograms of cell sequencing depth for raw gene expression data (left), after variance-stabilizing transform (middle), and after gene-wise scaling (right). **(B)** Selection of latent dimensionality through count splitting. The dashed line indicates k_{opt} selected by BacSC. **(C)** Selection of clustering resolution. The dashed line indicates the maximal gap in modularity between the training data Leiden clustering applied to the test data and a random clustering on the test data. **(D)** Selection of $n_{neighbors}$ and min_{dist} parameters for UMAP embedding with scDEED. The dashed black line shows the value of $n_{neighbors}$ selected by BacSC. **(E)** UMAP embedding generated with BacSC-selected parameters, colored by embedding reliability determined by scDEED with the same parameters. **(F)** Comparison of gene means of original and synthetic null data for DE testing. **(G)** Comparison of gene variances of original and synthetic null data for DE testing. **(H)** UMAP of synthetic null data for DE testing of each cell type against all other cells, colored by the two-group clustering determined for calculation of contrast scores. **(I)** Histograms of uncorrected p-values for DE testing of each cell type against all other cells. **(J)** Histograms of contrast scores for DE testing of each cell type against all other cells. The dashed lines indicate significance threshold values at the FDR level $\alpha = 0.05$.

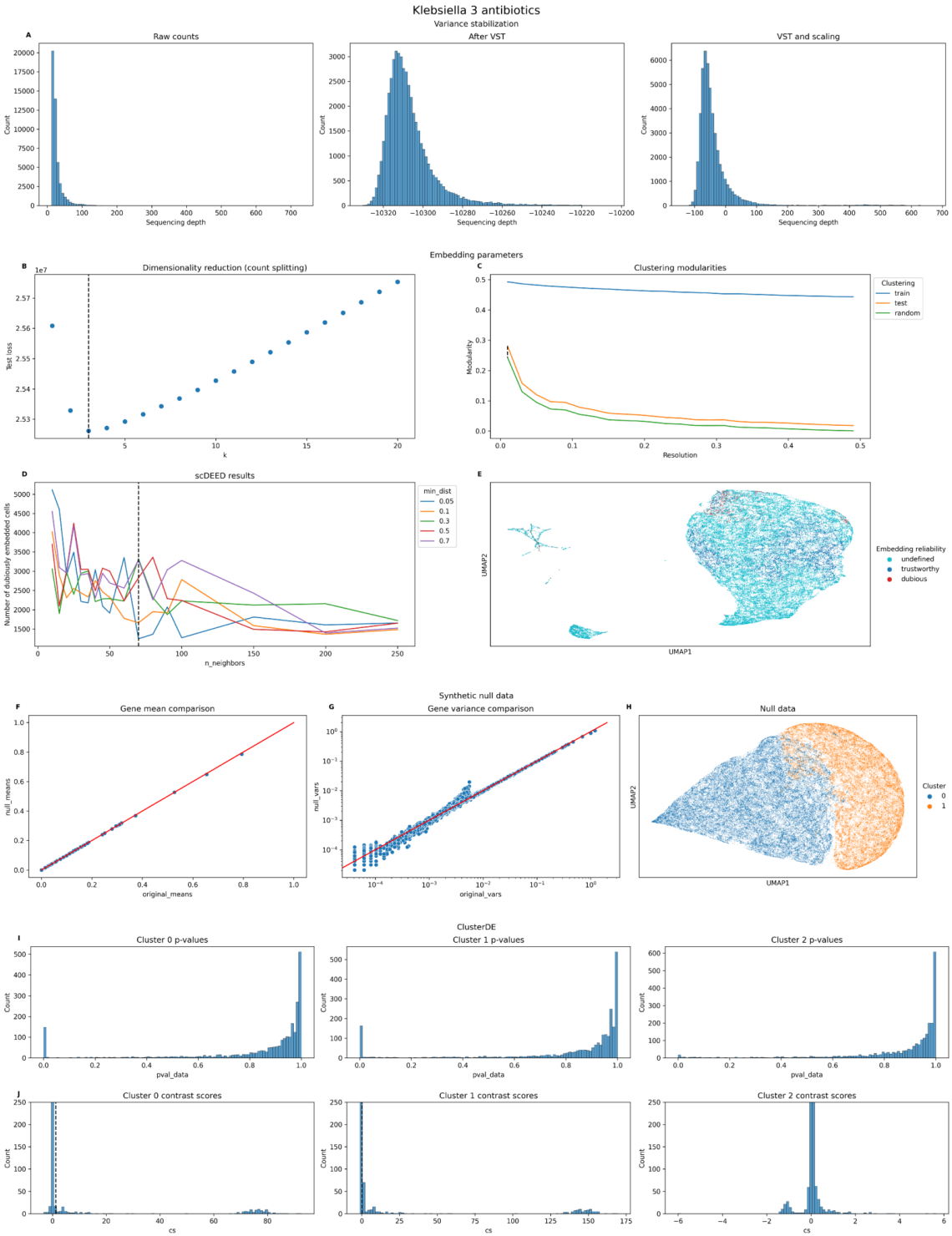


Fig. D21 Diagnostic plots generated during the analysis of the *Klebs_untreated_BD* dataset with BacSC. **(A)** Histograms of cell sequencing depth for raw gene expression data (left), after variance-stabilizing transform (middle), and after gene-wise scaling (right). **(B)** Selection of latent dimensionality through count splitting. The dashed line indicates k_{opt} selected by BacSC. **(C)** Selection of clustering resolution. The dashed line indicates the maximal gap in modularity between the training data Leiden clustering applied to the test data and a random clustering on the test data. **(D)** Selection of $n_{neighbors}$ and min_{dist} parameters for UMAP embedding with scDEED. The dashed black line shows the value of $n_{neighbors}$ selected by BacSC. **(E)** UMAP embedding generated with BacSC-selected parameters, colored by embedding reliability determined by scDEED with the same parameters. **(F)** Comparison of gene means of original and synthetic null data for DE testing. **(G)** Comparison of gene variances of original and synthetic null data for DE testing. **(H)** UMAP of synthetic null data for DE testing of each cell type against all other cells, colored by the two-group clustering determined for calculation of contrast scores. **(I)** Histograms of uncorrected p-values for DE testing of each cell type against all other cells. **(J)** Histograms of contrast scores for DE testing of each cell type against all other cells. The dashed lines indicate significance threshold values at the FDR level $\alpha = 0.05$.

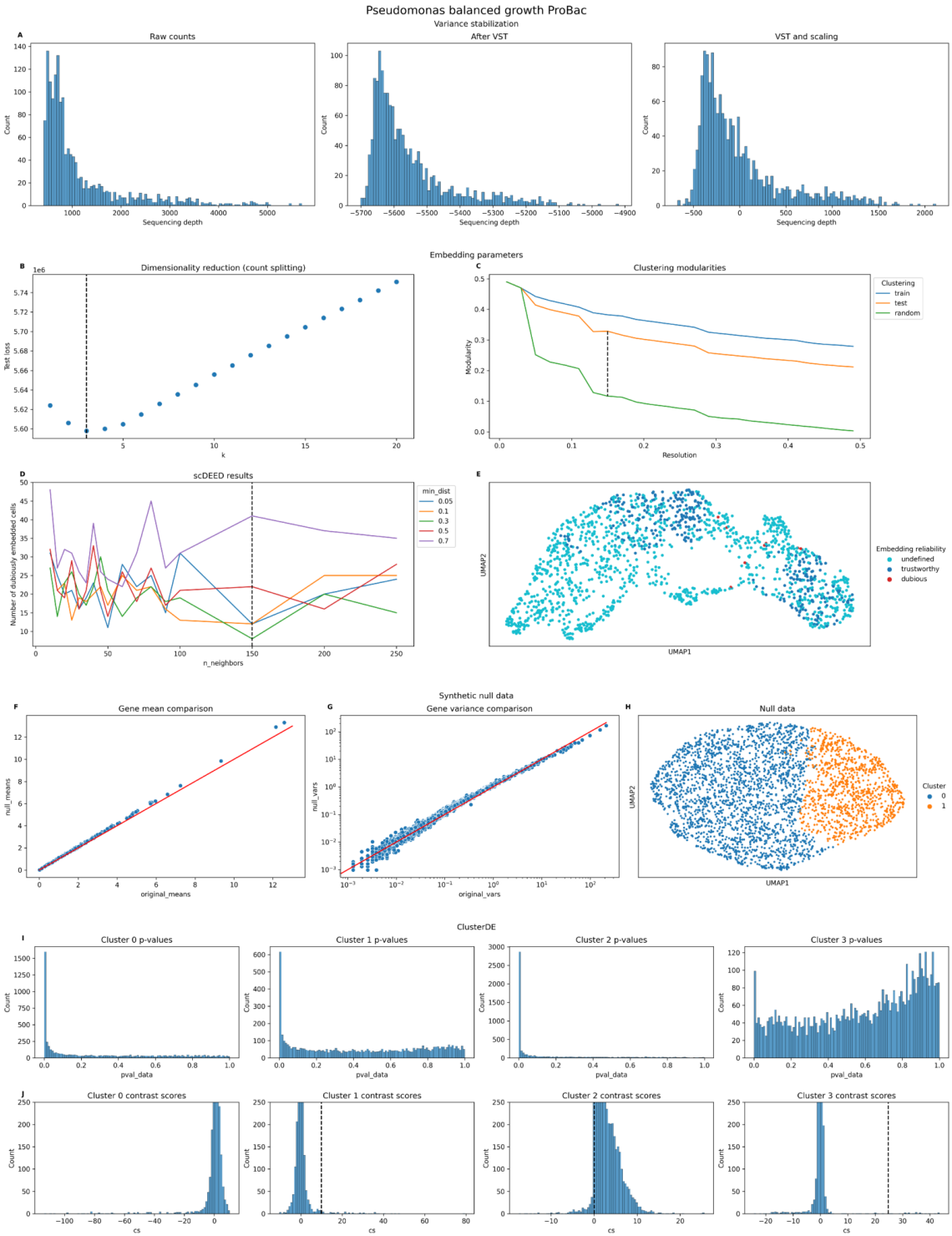


Fig. D22 Diagnostic plots generated during the analysis of the *Pseudomonas_balanced_PB* dataset with BacSC. (A) Histograms of cell sequencing depth for raw gene expression data (left), after variance-stabilizing transform (middle), and after gene-wise scaling (right). **(B)** Selection of latent dimensionality through count splitting. The dashed line indicates k_{opt} selected by BacSC. **(C)** Selection of clustering resolution. The dashed line indicates the maximal gap in modularity between the training data Leiden clustering applied to the test data and a random clustering on the test data. **(D)** Selection of $n_{neighbors}$ and min_{dist} parameters for UMAP embedding with scDEED. The dashed black line shows the value of $n_{neighbors}$ selected by BacSC. **(E)** UMAP embedding generated with BacSC-selected parameters, colored by embedding reliability determined by scDEED with the same parameters. **(F)** Comparison of gene means of original and synthetic null data for DE testing. **(G)** Comparison of gene variances of original and synthetic null data for DE testing. **(H)** UMAP of synthetic null data for DE testing of each cell type against all other cells, colored by the two-group clustering determined for calculation of contrast scores. **(I)** Histograms of uncorrected p-values for DE testing of each cell type against all other cells. **(J)** Histograms of contrast scores for DE testing of each cell type against all other cells. The dashed lines indicate significance threshold values at the FDR level $\alpha = 0.05$.

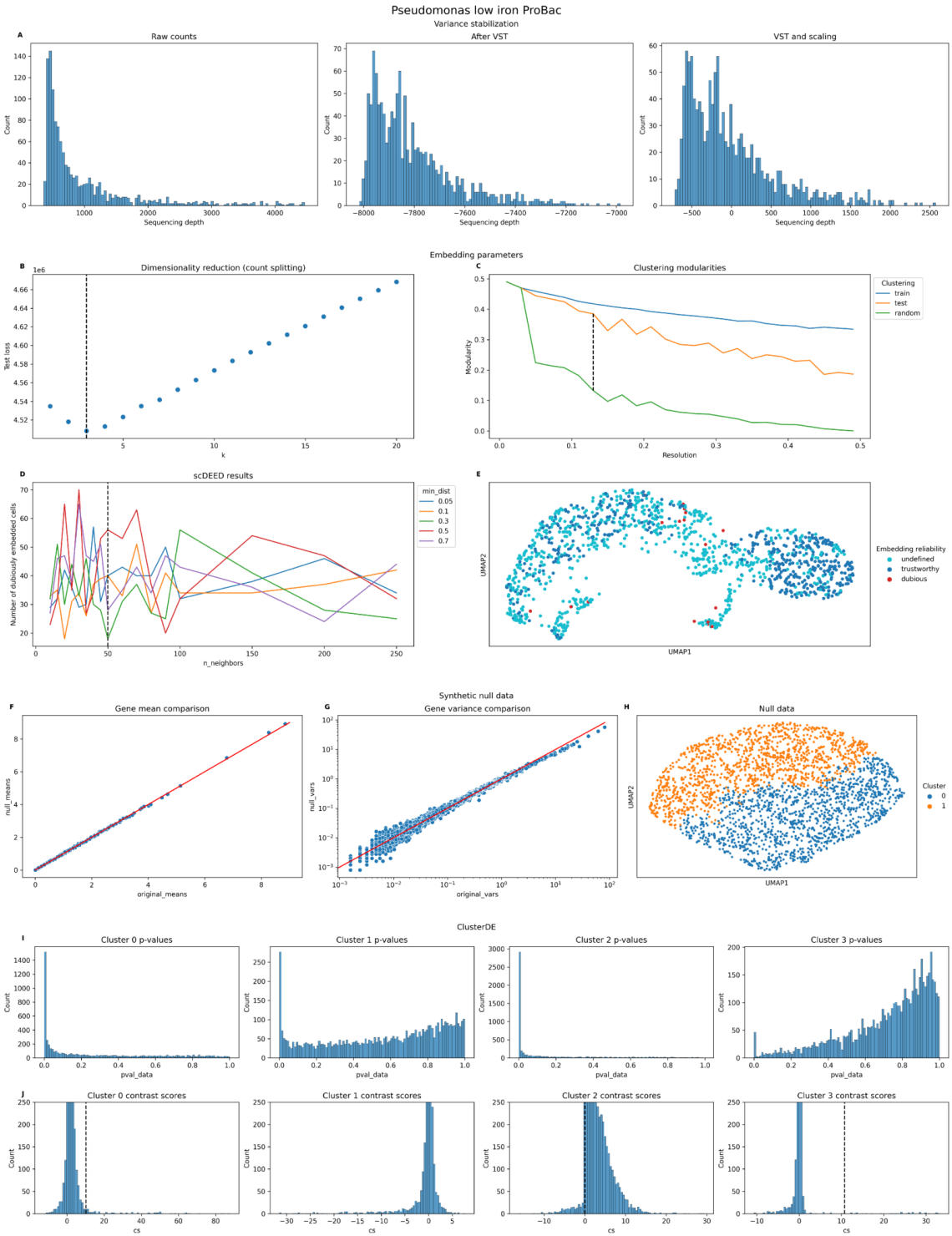


Fig. D23 Diagnostic plots generated during the analysis of the *Pseudomonas_li_PB* dataset with BacSC. **(A)** Histograms of cell sequencing depth for raw gene expression data (left), after variance-stabilizing transform (middle), and after gene-wise scaling (right). **(B)** Selection of latent dimensionality through count splitting. The dashed line indicates k_{opt} selected by BacSC. **(C)** Selection of clustering resolution. The dashed line indicates the maximal gap in modularity between the training data Leiden clustering applied to the test data and a random clustering on the test data. **(D)** Selection of $n_{neighbors}$ and min_{dist} parameters for UMAP embedding with scDEED. The dashed black line shows the value of $n_{neighbors}$ selected by BacSC. **(E)** UMAP embedding generated with BacSC-selected parameters, colored by embedding reliability determined by scDEED with the same parameters. **(F)** Comparison of gene means of original and synthetic null data for DE testing. **(G)** Comparison of gene variances of original and synthetic null data for DE testing. **(H)** UMAP of synthetic null data for DE testing of each cell type against all other cells, colored by the two-group clustering determined for calculation of contrast scores. **(I)** Histograms of uncorrected p-values for DE testing of each cell type against all other cells. **(J)** Histograms of contrast scores for DE testing of each cell type against all other cells. The dashed lines indicate significance threshold values at the FDR level $\alpha = 0.05$.

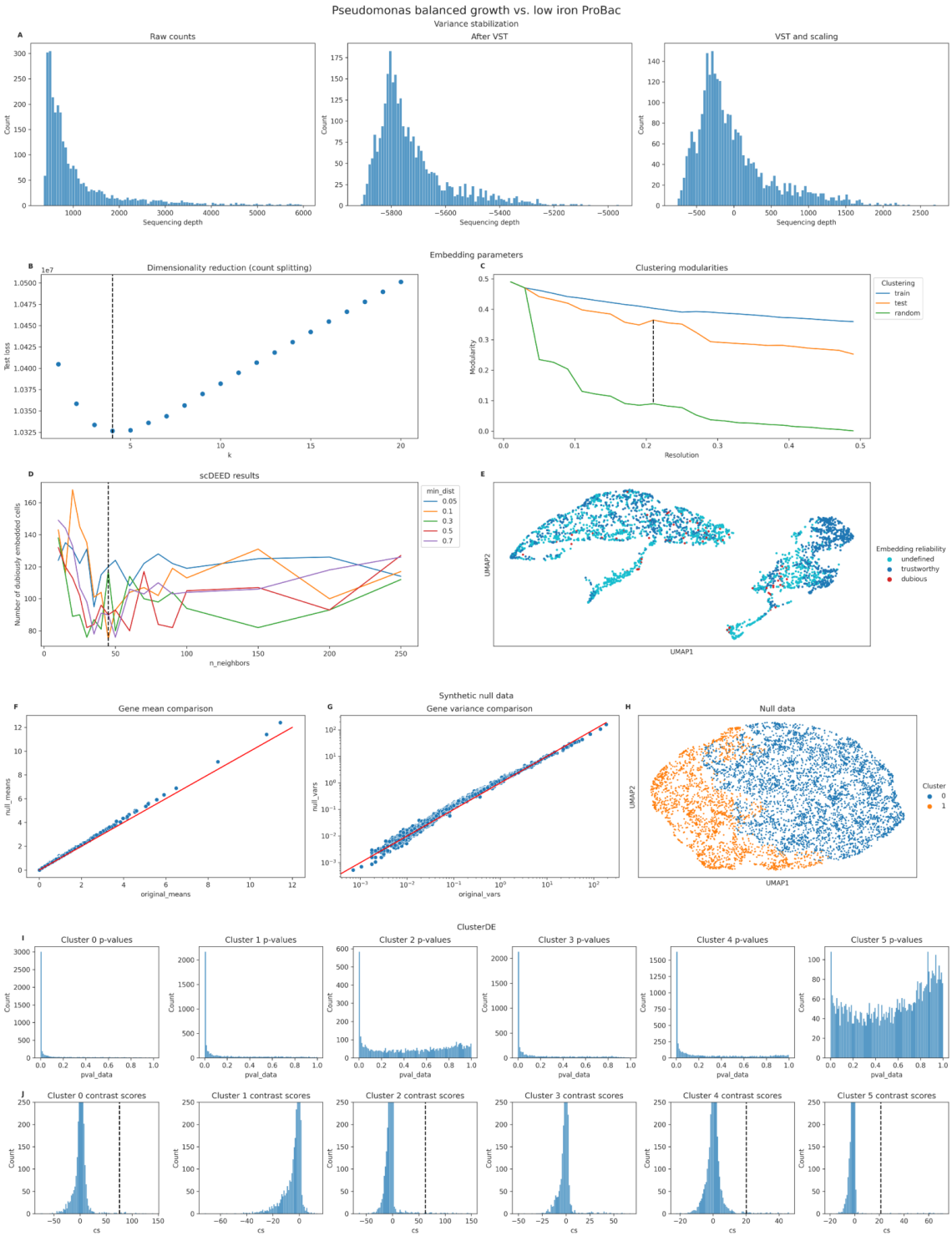


Fig. D24 Diagnostic plots generated during the analysis of the combined *Pseudomonas_balanced_PB* and *Pseudomonas_li_PB* dataset with BacSC. (A) Histograms of cell sequencing depth for raw gene expression data (left), after variance-stabilizing transform (middle), and after gene-wise scaling (right). **(B)** Selection of latent dimensionality through count splitting. The dashed line indicates k_{opt} selected by BacSC. **(C)** Selection of clustering resolution. The dashed line indicates the maximal gap in modularity between the training data Leiden clustering applied to the test data and a random clustering on the test data. **(D)** Selection of $n_{neighbors}$ and min_{dist} parameters for UMAP embedding with scDEED. The dashed black line shows the value of $n_{neighbors}$ selected by BacSC. **(E)** UMAP embedding generated with BacSC-selected parameters, colored by embedding reliability determined by scDEED with the same parameters. **(F)** Comparison of gene means of original and synthetic null data for DE testing. **(G)** Comparison of gene variances of original and synthetic null data for DE testing. **(H)** UMAP of synthetic null data for DE testing of each cell type against all other cells, colored by the two-group clustering determined for calculation of contrast scores. **(I)** Histograms of uncorrected p-values for DE testing of each cell type against all other cells. **(J)** Histograms of contrast scores for DE testing of each cell type against all other cells. The dashed lines indicate significance threshold values at the FDR level $\alpha = 0.05$.

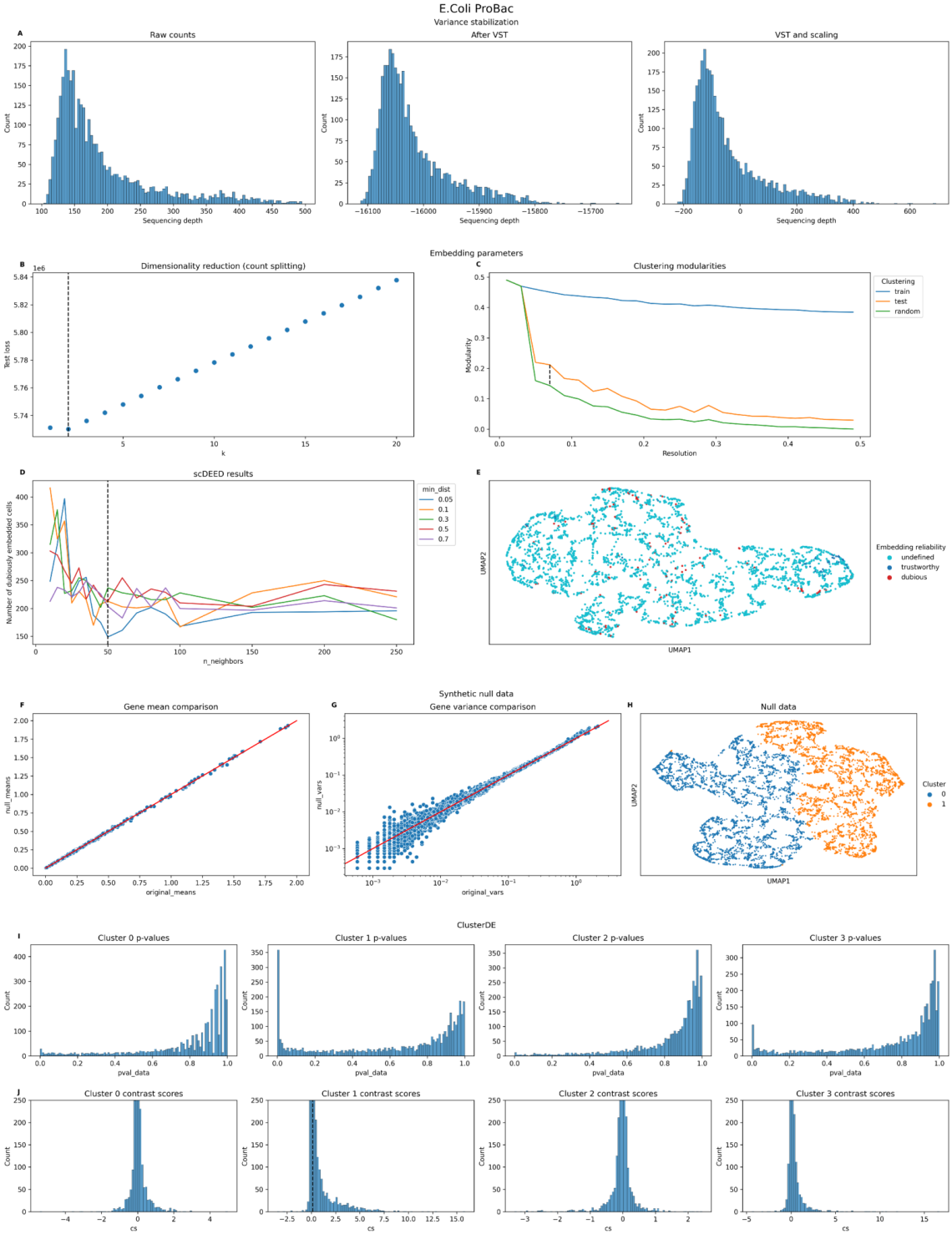


Fig. D25 Diagnostic plots generated during the analysis of the *Ecoli_balanced_PB* dataset with BacSC. (A) Histograms of cell sequencing depth for raw gene expression data (left), after variance-stabilizing transform (middle), and after gene-wise scaling (right). **(B)** Selection of latent dimensionality through count splitting. The dashed line indicates k_{opt} selected by BacSC. **(C)** Selection of clustering resolution. The dashed line indicates the maximal gap in modularity between the training data Leiden clustering applied to the test data and a random clustering on the test data. **(D)** Selection of $n_{neighbors}$ and min_{dist} parameters for UMAP embedding with scDEED. The dashed black line shows the value of $n_{neighbors}$ selected by BacSC. **(E)** UMAP embedding generated with BacSC-selected parameters, colored by embedding reliability determined by scDEED with the same parameters. **(F)** Comparison of gene means of original and synthetic null data for DE testing. **(G)** Comparison of gene variances of original and synthetic null data for DE testing. **(H)** UMAP of synthetic null data for DE testing of each cell type against all other cells, colored by the two-group clustering determined for calculation of contrast scores. **(I)** Histograms of uncorrected p-values for DE testing of each cell type against all other cells. **(J)** Histograms of contrast scores for DE testing of each cell type against all other cells. The dashed lines indicate significance threshold values at the FDR level $\alpha = 0.05$.

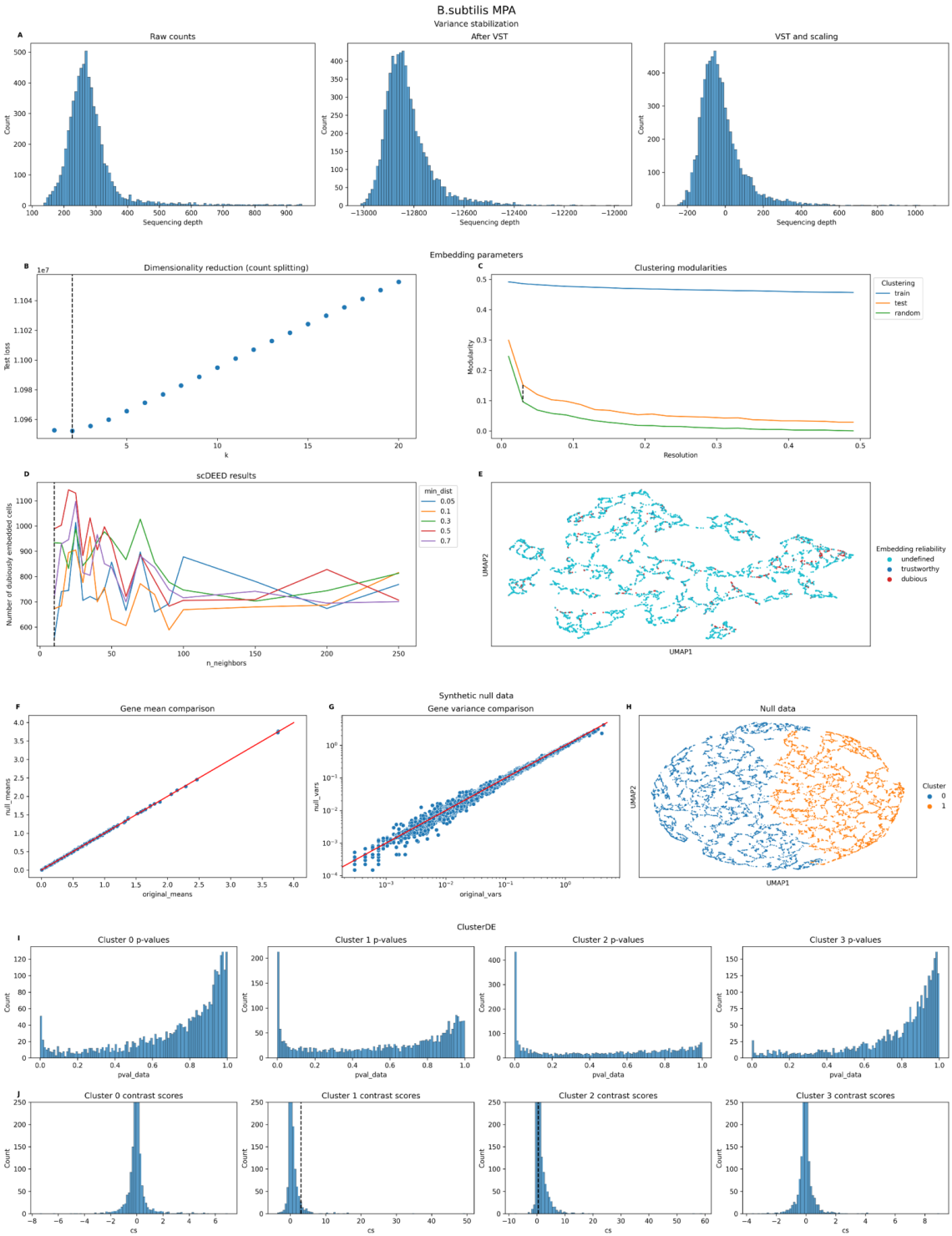


Fig. D26 Diagnostic plots generated during the analysis of the *Bsub_MPA_PB* dataset with BacSC. (A) Histograms of cell sequencing depth for raw gene expression data (left), after variance-stabilizing transform (middle), and after gene-wise scaling (right). **(B)** Selection of latent dimensionality through count splitting. The dashed line indicates k_{opt} selected by BacSC. **(C)** Selection of clustering resolution. The dashed line indicates the maximal gap in modularity between the training data Leiden clustering applied to the test data and a random clustering on the test data. **(D)** Selection of $n_{neighbors}$ and min_{dist} parameters for UMAP embedding with scDEED. The dashed black line shows the value of $n_{neighbors}$ selected by BacSC. **(E)** UMAP embedding generated with BacSC-selected parameters, colored by embedding reliability determined by scDEED with the same parameters. **(F)** Comparison of gene means of original and synthetic null data for DE testing. **(G)** Comparison of gene variances of original and synthetic null data for DE testing. **(H)** UMAP of synthetic null data for DE testing of each cell type against all other cells, colored by the two-group clustering determined for calculation of contrast scores. **(I)** Histograms of uncorrected p-values for DE testing of each cell type against all other cells. **(J)** Histograms of contrast scores for DE testing of each cell type against all other cells. The dashed lines indicate significance threshold values at the FDR level $\alpha = 0.05$.

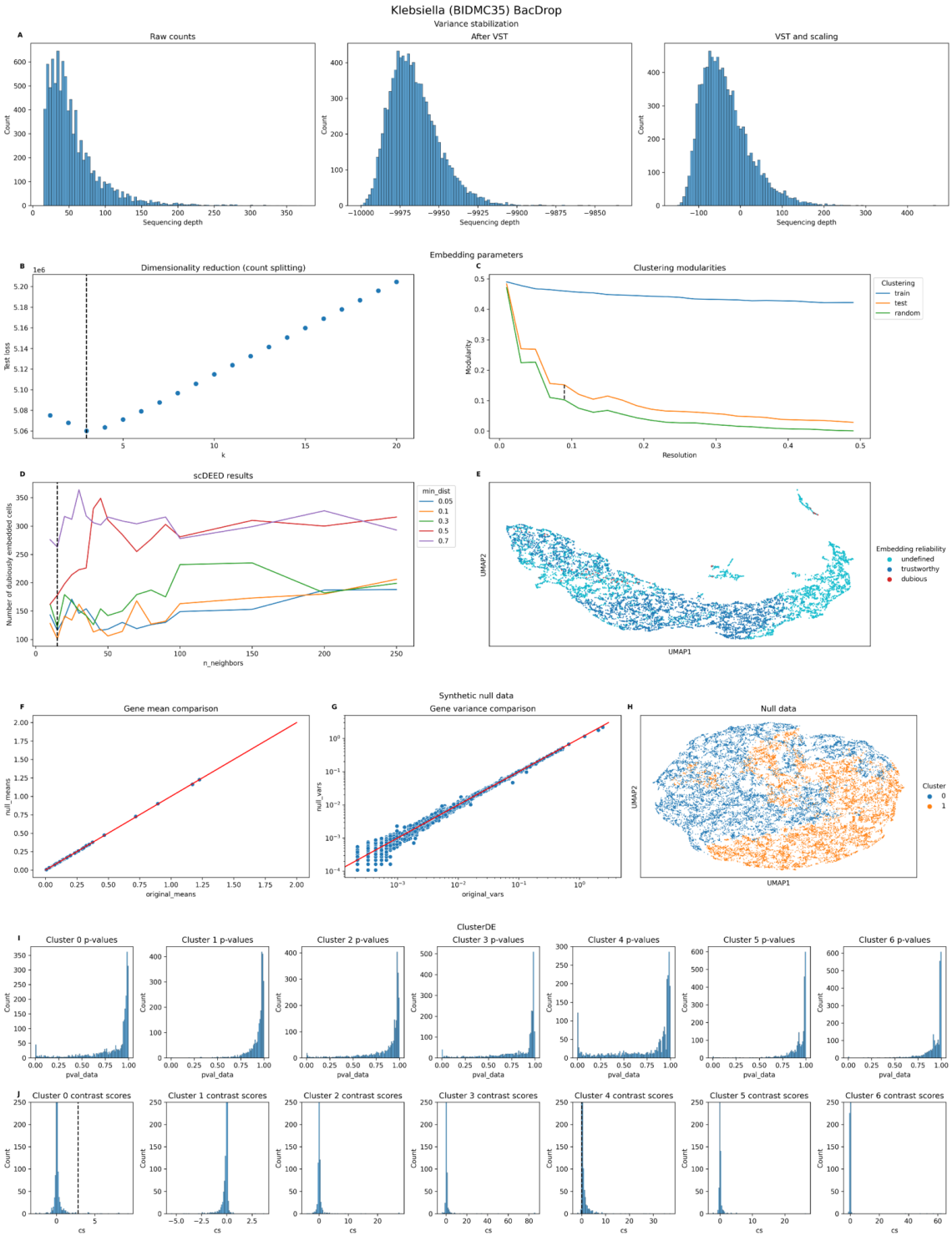


Fig. D27 Diagnostic plots generated during the analysis of the *Klebs_BIDMC35_BD* dataset with BacSC. **(A)** Histograms of cell sequencing depth for raw gene expression data (left), after variance-stabilizing transform (middle), and after gene-wise scaling (right). **(B)** Selection of latent dimensionality through count splitting. The dashed line indicates k_{opt} selected by BacSC. **(C)** Selection of clustering resolution. The dashed line indicates the maximal gap in modularity between the training data Leiden clustering applied to the test data and a random clustering on the test data. **(D)** Selection of $n_{neighbors}$ and min_{dist} parameters for UMAP embedding with scDEED. The dashed black line shows the value of $n_{neighbors}$ selected by BacSC. **(E)** UMAP embedding generated with BacSC-selected parameters, colored by embedding reliability determined by scDEED with the same parameters. **(F)** Comparison of gene means of original and synthetic null data for DE testing. **(G)** Comparison of gene variances of original and synthetic null data for DE testing. **(H)** UMAP of synthetic null data for DE testing of each cell type against all other cells, colored by the two-group clustering determined for calculation of contrast scores. **(I)** Histograms of uncorrected p-values for DE testing of each cell type against all other cells. **(J)** Histograms of contrast scores for DE testing of each cell type against all other cells. The dashed lines indicate significance threshold values at the FDR level $\alpha = 0.05$.

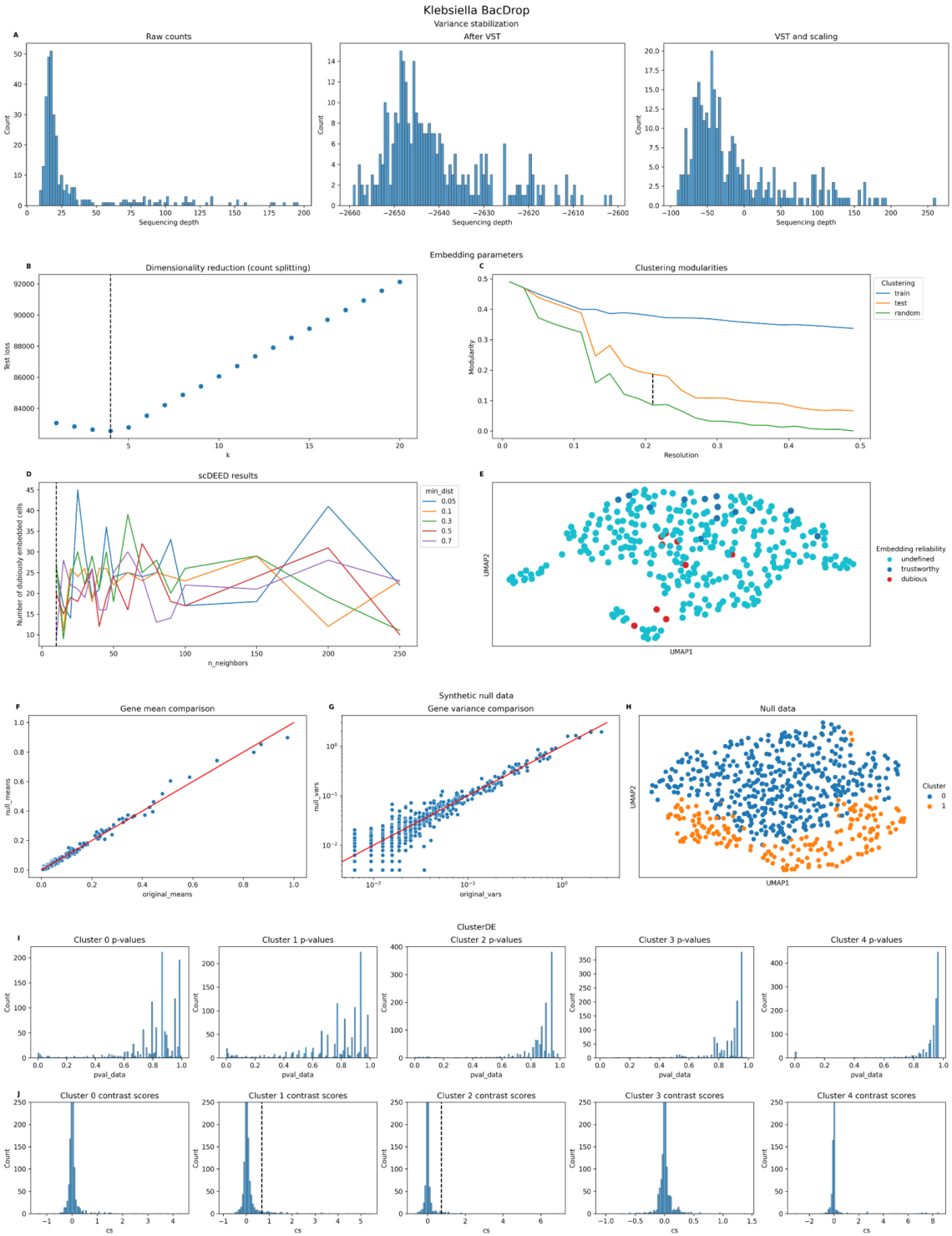


Fig. D28 Diagnostic plots generated during the analysis of the *Klebs_4species_BD* dataset with BacSC. (A) Histograms of cell sequencing depth for raw gene expression data (left), after variance-stabilizing transform (middle), and after gene-wise scaling (right). **(B)** Selection of latent dimensionality through count splitting. The dashed line indicates k_{opt} selected by BacSC. **(C)** Selection of clustering resolution. The dashed line indicates the maximal gap in modularity between the training data Leiden clustering applied to the test data and a random clustering on the test data. **(D)** Selection of $n_{neighbors}$ and min_{dist} parameters for UMAP embedding with scDEED. The dashed black line shows the value of $n_{neighbors}$ selected by BacSC. **(E)** UMAP embedding generated with BacSC-selected parameters, colored by embedding reliability determined by scDEED with the same parameters. **(F)** Comparison of gene means of original and synthetic null data for DE testing. **(G)** Comparison of gene variances of original and synthetic null data for DE testing. **(H)** UMAP of synthetic null data for DE testing of each cell type against all other cells, colored by the two-group clustering determined for calculation of contrast scores. **(I)** Histograms of uncorrected p-values for DE testing of each cell type against all other cells. **(J)** Histograms of contrast scores for DE testing of each cell type against all other cells. The dashed lines indicate significance threshold values at the FDR level $\alpha = 0.05$.

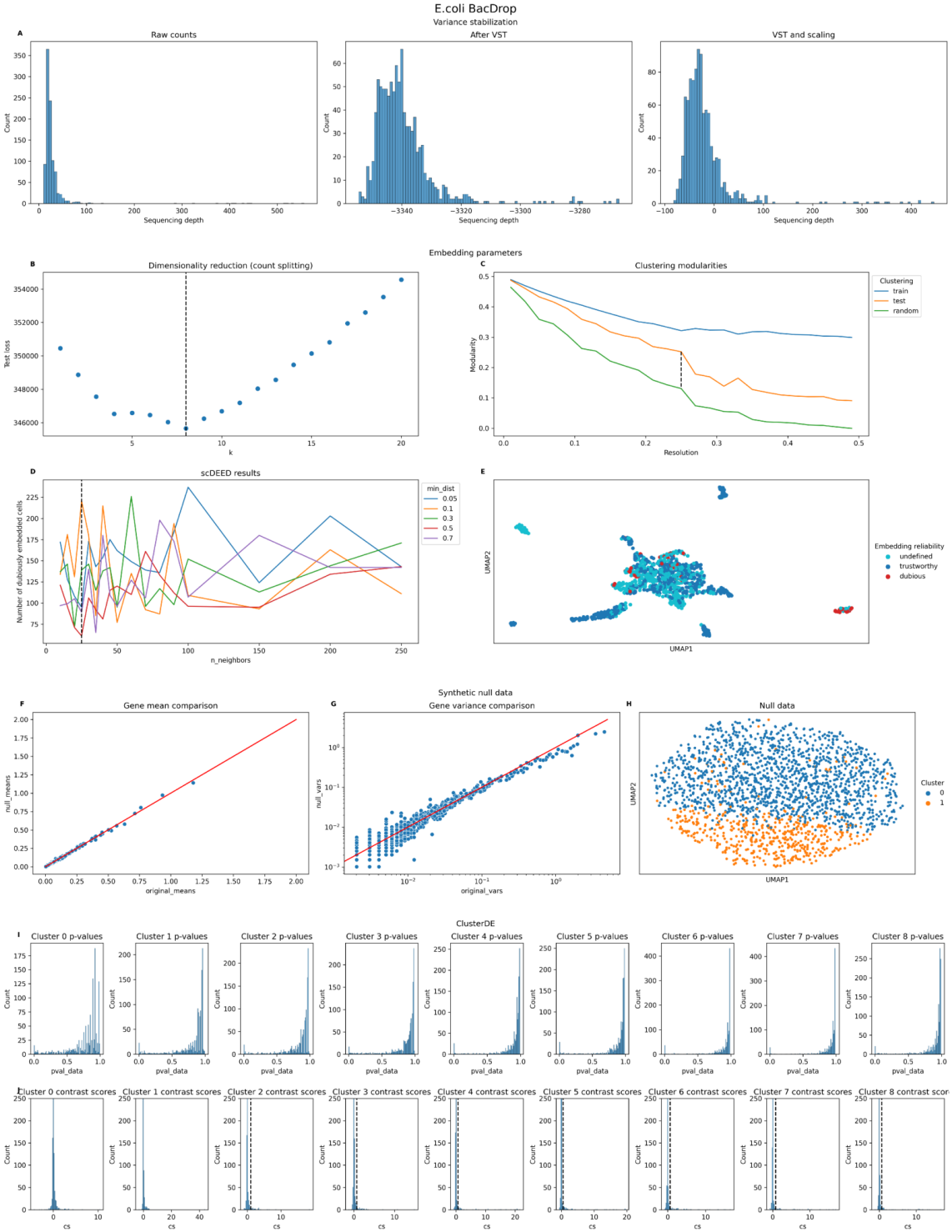


Fig. D29 Diagnostic plots generated during the analysis of the *Ecoli_4species_BD* dataset with BacSC. (A) Histograms of cell sequencing depth for raw gene expression data (left), after variance-stabilizing transform (middle), and after gene-wise scaling (right). **(B)** Selection of latent dimensionality through count splitting. The dashed line indicates k_{opt} selected by BacSC. **(C)** Selection of clustering resolution. The dashed line indicates the maximal gap in modularity between the training data Leiden clustering applied to the test data and a random clustering on the test data. **(D)** Selection of $n_{neighbors}$ and min_{dist} parameters for UMAP embedding with scDEED. The dashed black line shows the value of $n_{neighbors}$ selected by BacSC. **(E)** UMAP embedding generated with BacSC-selected parameters, colored by embedding reliability determined by scDEED with the same parameters. **(F)** Comparison of gene means of original and synthetic null data for DE testing. **(G)** Comparison of gene variances of original and synthetic null data for DE testing. **(H)** UMAP of synthetic null data for DE testing of each cell type against all other cells, colored by the two-group clustering determined for calculation of contrast scores. **(I)** Histograms of uncorrected p-values for DE testing of each cell type against all other cells. **(J)** Histograms of contrast scores for DE testing of each cell type against all other cells. The dashed lines indicate significance threshold values at the FDR level $\alpha = 0.05$.

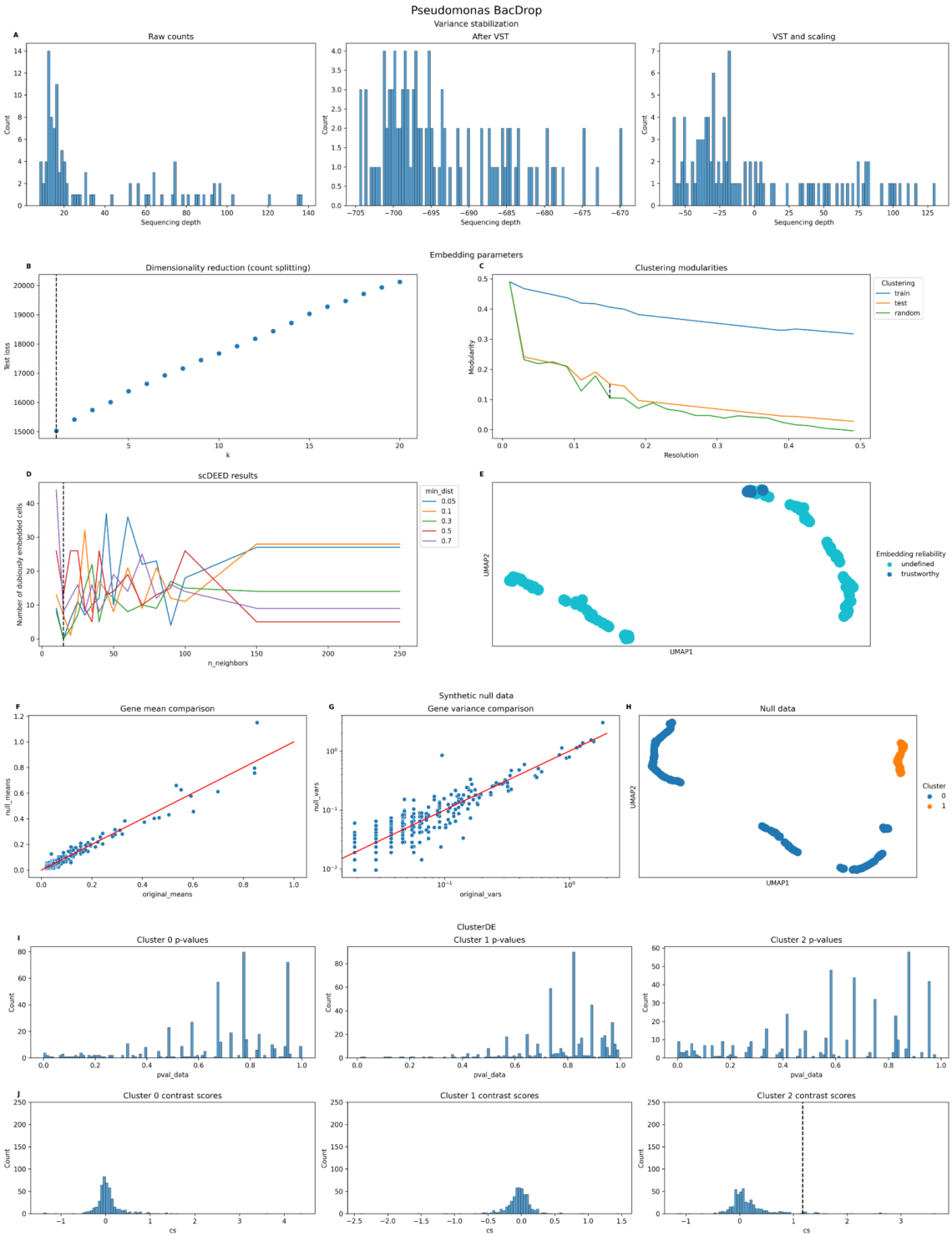


Fig. D30 Diagnostic plots generated during the analysis of the *Pseudomonas_4species_BD* dataset with BacSC. (A) Histograms of cell sequencing depth for raw gene expression data (left), after variance-stabilizing transform (middle), and after gene-wise scaling (right). **(B)** Selection of latent dimensionality through count splitting. The dashed line indicates k_{opt} selected by BacSC. **(C)** Selection of clustering resolution. The dashed line indicates the maximal gap in modularity between the training data Leiden clustering applied to the test data and a random clustering on the test data. **(D)** Selection of $n_{neighbors}$ and min_{dist} parameters for UMAP embedding with scDEED. The dashed black line shows the value of $n_{neighbors}$ selected by BacSC. **(E)** UMAP embedding generated with BacSC-selected parameters, colored by embedding reliability determined by scDEED with the same parameters. **(F)** Comparison of gene means of original and synthetic null data for DE testing. **(G)** Comparison of gene variances of original and synthetic null data for DE testing. **(H)** UMAP of synthetic null data for DE testing of each cell type against all other cells, colored by the two-group clustering determined for calculation of contrast scores. **(I)** Histograms of uncorrected p-values for DE testing of each cell type against all other cells. **(J)** Histograms of contrast scores for DE testing of each cell type against all other cells. The dashed lines indicate significance threshold values at the FDR level $\alpha = 0.05$.

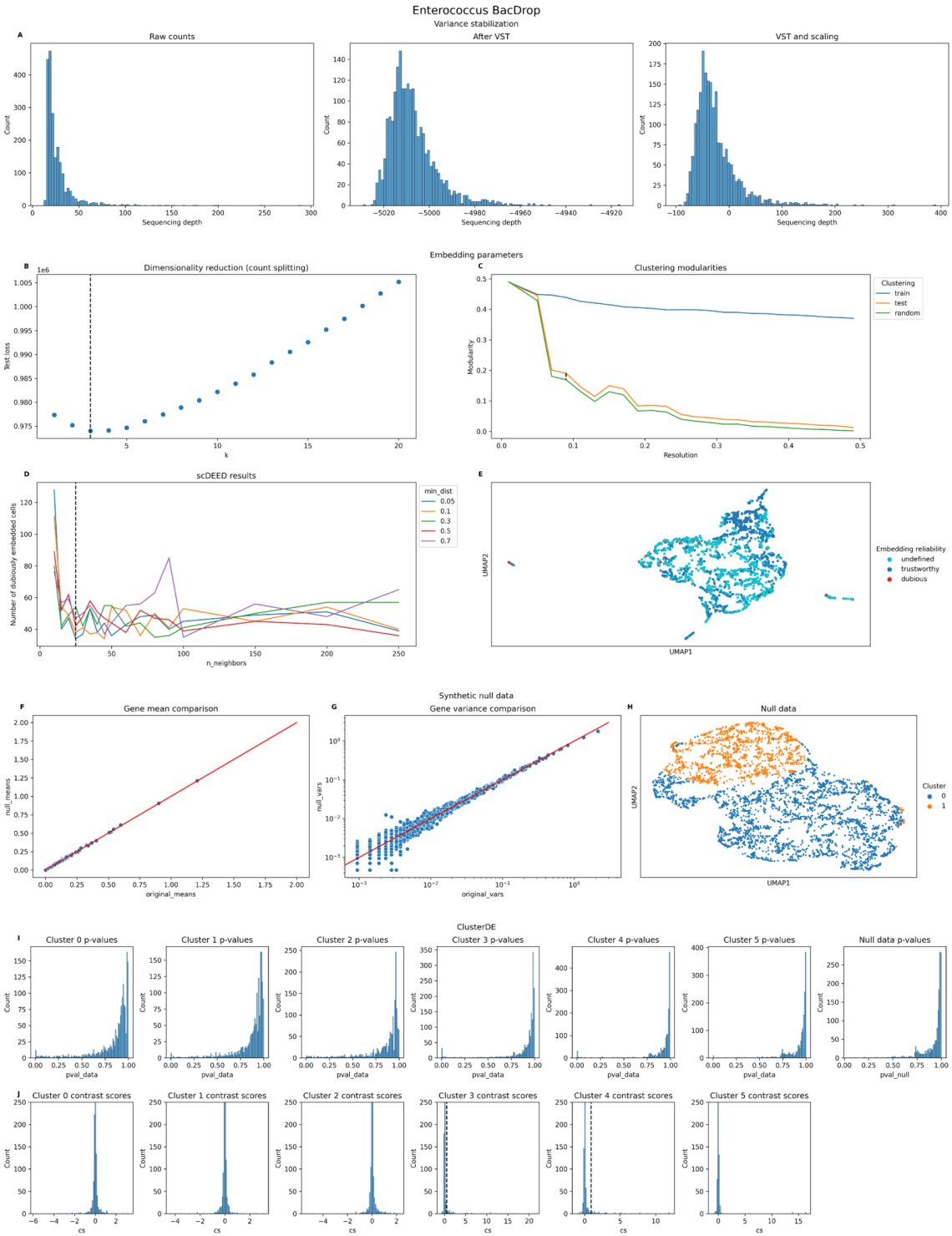


Fig. D31 Diagnostic plots generated during the analysis of the *Efaecium_4species_BD* dataset with BacSC. **(A)** Histograms of cell sequencing depth for raw gene expression data (left), after variance-stabilizing transform (middle), and after gene-wise scaling (right). **(B)** Selection of latent dimensionality through count splitting. The dashed line indicates k_{opt} selected by BacSC. **(C)** Selection of clustering resolution. The dashed line indicates the maximal gap in modularity between the training data Leiden clustering applied to the test data and a random clustering on the test data. **(D)** Selection of $n_{neighbors}$ and min_{dist} parameters for UMAP embedding with scDEED. The dashed black line shows the value of $n_{neighbors}$ selected by BacSC. **(E)** UMAP embedding generated with BacSC-selected parameters, colored by embedding reliability determined by scDEED with the same parameters. **(F)** Comparison of gene means of original and synthetic null data for DE testing. **(G)** Comparison of gene variances of original and synthetic null data for DE testing. **(H)** UMAP of synthetic null data for DE testing of each cell type against all other cells, colored by the two-group clustering determined for calculation of contrast scores. **(I)** Histograms of uncorrected p-values for DE testing of each cell type against all other cells. **(J)** Histograms of contrast scores for DE testing of each cell type against all other cells. The dashed lines indicate significance threshold values at the FDR level $\alpha = 0.05$.

Appendix E Supplementary tables

Dataset	Cells	Genes	Minimum seq. depth	Maximum seq. depth	Median seq. depth	Zero counts (percentage)	Maximum count	95% quantile	99% quantile
Pseudomonas_balanced_PB	1544	5553	413	5704	794.5	0.862	136.0	1.0	3.0
Pseudomonas_li_PB	1255	5540	360	4464	647.0	0.881	80.0	1.0	2.0
Ecoli_balanced_PB	3386	3968	103	495	163.0	0.963	14.0	0.0	1.0
Bsub_minmed_PB	2784	2952	141	1289	325.0	0.911	45.0	1.0	2.0
Bsub_damage_PB	13801	2959	268	1839	555.0	0.861	110.0	1.0	3.0
Bsub_MPA_PB	6703	2937	136	948	267.0	0.940	105.0	1.0	2.0
Klebs_anitbiotics_BD	19638	2500	14	275	21.0	0.992	13.0	0.0	0.0
Klebs_untreated_BD	48511	2500	12	728	21.0	0.991	30.0	0.0	0.0
Klebs_BIDMC35_BD	9168	2500	15	371	45.0	0.990	26.0	0.0	0.0
Klebs_4species_BD	315	1265	9	196	19.0	0.978	10.0	0.0	1.0
Ecoli_4species_BD	983	1301	10	556	21.0	0.981	35.0	0.0	1.0
Pseudomonas_4species_BD	103	628	8	137	18.0	0.953	7.0	0.0	1.0
Efaecium_4species_BD	2113	1606	12	289	22.0	0.985	19.0	0.0	1.0

Table E1 Dimensionality and summary statistics of datasets after quality control with BacSC. If not stated otherwise, statistics are in terms of counts/absolute values.

Dataset	Minimum sequencing depth	Minimum cells per gene	Median absolute deviation cutoff (*nmads*)	Number of removed barcodes	Data distribution	Latent dimension (k_{opt})	n_neighbors	min_dist	clustering resolution
Pseudomonas_balanced_PB	-	2	5	108	NB	3	150	0.30	0.15
Pseudomonas_li_PB	-	2	5	71	NB	3	50	0.30	0.13
Ecoli_balanced_PB	100	2	5	1376	Poi	2	50	0.05	0.07
Bsub_minmed_PB	100	2	5	0	Poi	4	20	0.50	0.15
Bsub_damage_PB	100	2	5	61	Poi	8	150	0.30	0.37
Bsub_sporulation_PB	50	2	30	10204	Poi	4	250	0.30	0.29
Bsub_MPA_PB	100	2	10	197	Poi	2	10	0.05	0.03
Klebs_anitbiotics_BD	15	2	15	1214846	Poi	5	150	0.10	0.17
Klebs_untreated_BD	15	2	15	409547	Poi	3	70	0.05	0.01
Klebs_BIDMC35_BD	15	2	5	768	Poi	3	15	0.10	0.09
Klebs_4species_BD	15	2	10	8335	Poi	4	10	0.70	0.21
Ecoli_4species_BD	15	2	10	8671	NB	7	25	0.50	0.25
Pseudomonas_4species_BD	15	2	10	8089	Poi	1	15	0.05	0.15
Efaecium_4species_BD	15	2	10	7862	Poi	3	25	0.05	0.09

Table E2 Overview over filtering thresholds used for quality control, number of removed barcodes, and hyperparameters determined during the course of BacSC in each dataset. Both *P.aeruginosa* datasets generated with ProBac-seq were already quality-controlled in CellRanger and therefore needed no further cell filtering for minimal sequencing depth. The "Data distribution" column denotes the data distribution determined for count splitting (see Methods). "NB" stands for the Negative Binomial distribution, "Poi" denotes the Poisson distribution.

Gene	Symbol	Name	PGFam	Rank (Wilcoxon test)
PA4514	NaN	iron transport outer membrane receptor	NaN	1
PA4370	icmP	insulin-cleaving metalloproteinase outer membrane protein	NaN	2
PA4515	NaN	hydroxylase	NaN	4
PA5531	tonB1	transporter TonB	NaN	6
PA4709	NaN	hemin degrading factor	NaN	9
PA4710	phuR	heme/hemoglobin uptake outer membrane receptor PhuR	NaN	10
PA4516	NaN	hypothetical protein	NaN	11
PA4707	NaN	ABC transporter permease	NaN	13
PA0472	NaN	RNA polymerase sigma factor	RNA polymerase ECF-type sigma factor	14
PA0672	hemO	heme oxygenase	Heme oxygenase HemO, associated with heme uptake	16
PA2468	foxI	ECF sigma factor FoxI	FIG006045: Sigma factor, ECF subfamily	17
PA2426	pvdS	extracytoplasmic-function sigma-70 factor	Sigma factor PvdS, controling pyoverdin biosynthesis	18
PA4371	NaN	hypothetical protein	NaN	19
PA4513	NaN	oxidoreductase	NaN	20
PA0929	NaN	two-component response regulator	Two-component transcriptional response regulator, LuxR family	21
PA2467	foxR	anti-sigma factor FoxR	Iron siderophore sensor protein	24
PA4468	sodM	superoxide dismutase	NaN	26
PA3530	NaN	hypothetical protein	NaN	28
PA0931	pirA	outer membrane receptor FepA	TonB-dependent receptor; Outer membrane receptor for ferric enterobactin and colicins B, D	31
PA5217	NaN	iron ABC transporter substrate-binding protein	NaN	34
PA3899	NaN	RNA polymerase sigma factor	NaN	36
PA4470	fumC1	fumarate hydratase	NaN	39
PA4708	phuT	heme-transporter PhuT	NaN	40
PA4227	pchR	transcriptional regulator PchR	NaN	42
PA1911	femR	sigma factor regulator FemR	Iron siderophore sensor protein	43
PA4168	fpvB	second ferric pyoverdine receptor FpvB	NaN	45
PA0930	NaN	two-component sensor	two-component sensor	55
PA1912	femI	ECF sigma factor FemI	FIG006045: Sigma factor, ECF subfamily	59
PA3900	NaN	transmembrane sensor	NaN	71
PA1300	NaN	ECF subfamily sigma-70 factor	FIG006045: Sigma factor, ECF subfamily	73
PA0471	NaN	transmembrane sensor	Putative transmembrane sensor	79
PA4706	NaN	hemin importer ATP-binding subunit	NaN	81
PA2033	NaN	hypothetical protein	Siderophore-interacting protein	86
PA1365	NaN	siderophore receptor	Ferrichrome-iron receptor @ Iron siderophore receptor protein	99
PA4471	NaN	hypothetical protein	NaN	105
PA4705	NaN	hypothetical protein	NaN	108
PA1802	clpX	ATP-dependent protease ATP-binding subunit ClpX	ATP-dependent Clp protease ATP-binding subunit ClpX	113
PA1301	NaN	transmembrane sensor	Iron siderophore sensor protein	137
PA4467	NaN	hypothetical protein	NaN	153
PA0800	NaN	hypothetical protein	FIG024006: iron uptake protein	154
PA4469	NaN	hypothetical protein	NaN	155
PA5148	NaN	hypothetical protein	NaN	158

Table E3 Description of genes and rank of p-value from DE testing balanced growth versus low-iron in the combined *Pseudomonas_balanced_PB* and *Pseudomonas_li_PB* dataset. Only genes that are DE in the Copathogenex dataset for at least one of the three DE tests performed on that data are shown

Cell Type	Number of cells	minimal q	DE genes, FDR = 0.05	DE genes, FDR = 0.1	DE genes, FDR = 0.2
0	1219	0.059649	0	1804	2195
1	897	0.148019	0	0	1517
2	386	0.257908	0	0	0
3	262	0.027027	47	50	50
4	20	0.035714	28	34	62

Table E4 Description of clusters for the *Bsub_minmed_PB* dataset. The table shows number of cells, minimal FDR (q value) over all genes, and number of differentially expressed genes at three different FDR levels.

Cell Type	Number of cells	minimal q	DE genes, FDR = 0.05	DE genes, FDR = 0.1	DE genes, FDR = 0.2
0	7954	0.076923	0	13	25
1	5960	0.102564	0	0	111
2	3262	0.052632	0	62	667
3	1843	0.016129	96	122	673
4	255	0.111111	0	0	41
5	223	0.029412	69	83	113
6	74	0.016667	102	121	160
7	67	0.012987	102	118	133

Table E5 Description of clusters for the *Klebs_antibiotics_BD* dataset. The table shows number of cells, minimal FDR (q value) over all genes, and number of differentially expressed genes at three different FDR levels.

Cell Type	Number of cells	minimal q	DE genes, FDR = 0.05	DE genes, FDR = 0.1	DE genes, FDR = 0.2
0	571	0.273290	0	0	0
1	484	0.020833	51	71	81
2	415	0.028825	5056	5209	5209
3	74	0.045455	22	23	27

Table E6 Description of clusters for the *Pseudomonas_balanced_PB* dataset. The table shows number of cells, minimal FDR (q value) over all genes, and number of differentially expressed genes at three different FDR levels.

Cell Type	Number of cells	minimal q	DE genes, FDR = 0.05	DE genes, FDR = 0.1	DE genes, FDR = 0.2
0	516	0.014925	82	4462	4850
1	446	0.798621	0	0	0
2	239	0.030303	5105	5210	5210
3	54	0.029412	34	35	36

Table E7 Description of clusters for the *Pseudomonas_li_PB* dataset. The table shows number of cells, minimal FDR (q value) over all genes, and number of differentially expressed genes at three different FDR levels.

Cell Type	Number of cells	minimal q	DE genes, FDR = 0.05	DE genes, FDR = 0.1	DE genes, FDR = 0.2
0	777	0.041667	24	31	44
1	773	1.000000	0	0	0
2	576	0.025000	43	54	66
3	396	0.120000	0	0	34
4	194	0.025000	50	66	71
5	124	0.029412	34	36	36

Table E8 Description of clusters for the combined *Pseudomonas_balanced_PB* and *Pseudomonas_li_PB* dataset. The table shows number of cells, minimal FDR (q value) over all genes, and number of differentially expressed genes at three different FDR levels.

Cell Type	Number of cells	minimal q	DE genes, FDR = 0.05	DE genes, FDR = 0.1	DE genes, FDR = 0.2
0	1132	0.416667	0	0	0
1	796	0.006452	1423	1821	2374
2	729	1.000000	0	0	0
3	729	0.055556	0	281	562

Table E9 Description of clusters for the *Ecoli_balanced_PB* dataset. The table shows number of cells, minimal FDR (q value) over all genes, and number of differentially expressed genes at three different FDR levels.

Cell Type	Number of cells	minimal q	DE genes, FDR = 0.05	DE genes, FDR = 0.1	DE genes, FDR = 0.2
0	5166	0.263282	0	0	0
1	3734	0.086339	0	2694	2694
2	3246	0.083924	0	2049	2467
3	576	1.000000	0	0	0
4	526	0.778626	0	0	0
5	422	0.500000	0	0	0
6	131	0.100000	0	0	11

Table E10 Description of clusters for the *Bsub_damage_PB* dataset. The table shows number of cells, minimal FDR (q value) over all genes, and number of differentially expressed genes at three different FDR levels.

Cell Type	Number of cells	minimal q	DE genes, FDR = 0.05	DE genes, FDR = 0.1	DE genes, FDR = 0.2
0	2275	0.388889	0	0	0
1	1940	0.019231	66	245	649
2	1602	0.008163	926	1634	2158
3	886	0.200000	0	0	0

Table E11 Description of clusters for the *Bsub_MPA_PB* dataset. The table shows number of cells, minimal FDR (q value) over all genes, and number of differentially expressed genes at three different FDR levels.

Cell Type	Number of cells	minimal q	DE genes, FDR = 0.05	DE genes, FDR = 0.1	DE genes, FDR = 0.2
0	44236	0.007299	148	161	362
1	2194	0.005988	324	412	676
2	2081	0.095238	0	21	21

Table E12 Description of clusters for the *Klebs_untreated_BD* dataset. The table shows number of cells, minimal FDR (q value) over all genes, and number of differentially expressed genes at three different FDR levels.

Cell Type	Number of cells	minimal q	DE genes, FDR = 0.05	DE genes, FDR = 0.1	DE genes, FDR = 0.2
0	2504	0.050000	0	20	26
1	1892	1.000000	0	0	0
2	1807	0.125000	0	0	8
3	1589	0.066667	0	15	86
4	914	0.008696	1047	1237	1561
5	255	0.142857	0	0	41
6	207	0.142857	0	0	7

Table E13 Description of clusters for the *Klebs_BIDMC35_BD* dataset. The table shows number of cells, minimal FDR (q value) over all genes, and number of differentially expressed genes at three different FDR levels.

Cell Type	Number of cells	minimal q	DE genes, FDR = 0.05	DE genes, FDR = 0.1	DE genes, FDR = 0.2
0	137	0.071429	0	14	53
1	96	0.019608	62	86	111
2	42	0.041667	24	31	36
3	26	0.333333	0	0	0
4	14	0.062500	0	29	30

Table E14 Description of clusters for the *Klebs_4species_PB* dataset. The table shows number of cells, minimal FDR (q value) over all genes, and number of differentially expressed genes at three different FDR levels.

Cell Type	Number of cells	minimal q	DE genes, FDR = 0.05	DE genes, FDR = 0.1	DE genes, FDR = 0.2
0	592	0.258621	0	0	0
1	107	0.062500	0	97	148
2	83	0.030303	33	41	82
3	56	0.040000	43	72	84
4	33	0.025641	39	58	69
5	30	0.018519	56	62	72
6	29	0.021277	52	54	63
7	28	0.027027	37	37	53
8	25	0.043478	43	47	53

Table E15 Description of clusters for the *Ecoli_4species_PB* dataset. The table shows number of cells, minimal FDR (q value) over all genes, and number of differentially expressed genes at three different FDR levels.

Cell Type	Number of cells	minimal q	DE genes, FDR = 0.05	DE genes, FDR = 0.1	DE genes, FDR = 0.2
0	42	0.206897	0	0	0
1	32	1.000000	0	0	0
2	29	0.043478	23	65	144

Table E16 Description of clusters for the *Pseudomonas_4species_PB* dataset. The table shows number of cells, minimal FDR (q value) over all genes, and number of differentially expressed genes at three different FDR levels.

Cell Type	Number of cells	minimal q	DE genes, FDR = 0.05	DE genes, FDR = 0.1	DE genes, FDR = 0.2
0	943	0.755102	0	0	0
1	589	1.000000	0	0	0
2	488	0.571429	0	0	0
3	36	0.018868	63	73	99
4	33	0.022727	48	53	89
5	24	0.100000	0	0	11

Table E17 Description of clusters for the *Efaecium_4species_PB* dataset. The table shows number of cells, minimal FDR (q value) over all genes, and number of differentially expressed genes at three different FDR levels.

C.2. Best practices for single-cell analysis across modalities

Contributing article

Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., [...] **Single-cell best practices consortium***, Schiller, H.B., Theis, F.J. (2023). Best practices for single-cell analysis across modalities. *Nat. Rev. Genet. 24, 550–572.* doi: `https://doi.org/10.1038/s41576-023-00586-w`

* I am part of the Single-cell best practices consortium

Additional information

Single-Cell Best Practices online book: `https://sc-best-practices.org`.

Copyright information

Author contributions

Main author list: A.C.S., L. Heumos and F.J.T. conceived the project. L. Heumos and A.C.S. contributed equally and have the right to list their name first in their curriculum vitae. A.C.S., L. Heumos, C.L. and F.D. wrote the manuscript. L.Z. and M.D.L. provided expertise for the discussion on transcriptomics; C.L. on chromatin accessibility; D.C.S. on surface protein expression; F.D., J.H. and F.C. on adaptive immune receptor repertoire analysis; and A.L. and F.C. on multimodal data integration. F.J.T. and H.B.S. supervised the work. Single-cell Best Practices Consortium: A.F., H.A., I.L.I., L.D., L.S., M.B., M.L., P.W., S.H.-z., Z.P., M.G.J., A.S., H.S., D.H., E.D., J.O., I.V., D.D., R.P., C.L.M., J.S.-R., J.H., P.B.M. and M.N. provided expertise for the discussion on transcriptomics; L.D.M. and I.L.I. on chromatin accessibility; C.R.-S. on surface protein expression; B.S. on adaptive immune receptor repertoire analysis; and G.P., L. Hetzel, J.T. and J.S.-R. on single-cell data resolved in space. M.A. contributed to the figure design. All authors read, edited and approved the final manuscript.

Eidesstattliche Versicherung (Affidavit)

(Siehe Promotionsordnung vom 12. Juli 2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 09.12.2024

Johannes Ostner