STATISTICAL METHODS FOR TRANSCRIPTOMIC DECONVOLUTION AND TEMPORAL GENE EXPRESSION MODELING

by

Suxian Zhou

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Applied Mathematics

Charlotte

2026

Approved by:

_____

Dr. Shaoyu Li

_____

Dr. Duan Chen

_____

Dr. Wenyu Gao

_____

Dr. Kirill Afonin

ABSTRACT

SUXIAN ZHOU. Statistical Methods for Transcriptomic Deconvolution and Temporal Gene Expression Modeling. (Under the direction of DR. SHAOYU LI)

High-throughput RNA sequencing technologies have transformed the study of cellular heterogeneity and dynamic gene expression programs. Bulk RNA sequencing provides aggregate measurements from mixed cell populations, whereas single-cell RNA sequencing offers cell-level resolution but is often affected by sparsity, technical noise, and limited sample coverage. These complementary characteristics motivate the development of statistical methods and computational algorithms that integrate bulk and single-cell transcriptomic data to estimate latent cellular compositions and characterize temporal gene expression dynamics.

This dissertation develops two statistical learning frameworks for transcriptomic data analysis. The first framework, GSNMF+, addresses bulk RNA-seq deconvolution by incorporating single-cell reference information into a geometry-guided nonnegative matrix factorization model. Standard NMF-based deconvolution is often ill-posed and sensitive to initialization and noise in the data. To improve robustness and interpretability, GSNMF+ introduces augmented pseudo-bulk mixtures, a solvability-guided regularization term, and a manifold-based penalty to encourage biologically meaningful latent components and stable proportion estimates. Simulation studies with known ground-truth proportions demonstrate that GSNMF+ improves deconvolution accuracy compared with existing approaches. Real bulk RNA-seq applications further show that the method produces more consistent stage-composition estimates across independent single-cell reference datasets.

The second framework, BetaDE, focuses on pseudotime-based differential expression analysis for single-cell RNA-seq data. Instead of relying on a single smooth trajectory model, BetaDE represents temporal gene expression patterns using a col-

lection of beta-shaped basis functions that capture diverse activation patterns along pseudotime. To accommodate the distributional features of single-cell count data, including overdispersion and excess zeros, BetaDE considers multiple count-based models, including Poisson, Negative Binomial, zero-inflated Poisson, and zero-inflated Negative Binomial models. Model and kernel selection are performed using Akaike Information Criterion, followed by hypothesis testing to identify genes with significant temporal expression changes. The fitted kernel-based features are further used for functional clustering to recover groups of genes with similar dynamic expression patterns.

Together, these two projects address complementary challenges in transcriptomic analysis: estimating hidden cellular compositions from bulk RNA-seq data and modeling dynamic gene expression programs from single-cell RNA-seq data. By combining matrix factorization, geometric regularization, data augmentation, pseudotime modeling, flexible count distributions, and functional clustering, this dissertation provides statistical tools for integrative analysis of bulk and single-cell RNA-seq data, with applications demonstrating their utility in complex biological systems.

DEDICATION

To my parents, Li Zhang and Rui Zhou, and to my grandparents, Hongying Wu and Tieniu Zhang.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Shaoyu Li, for her continuous guidance, support, and encouragement throughout my Ph.D. study. Her patience in addressing my questions, her insightful advice on my academic work, and her constant encouragement have been invaluable to my development as a researcher and educator. I am especially grateful for the many opportunities she has provided, which have greatly contributed to my professional growth. Beyond academic guidance, Dr. Li has also shown genuine care for my personal well-being and has provided tremendous support during my job search process. Her timely advice, encouragement, and generous assistance have meant a great deal to me.

I would also like to extend my sincere gratitude to my committee members, Dr. Duan Chen, Dr. Wenyu Gao, and Dr. Kirill Afonin, for their flexibility, support, and valuable feedback. I am also grateful to the faculty members in the department who have supported me throughout my Ph.D. journey. In particular, I would like to thank Dr. Shaozhong Deng for his continuous guidance, encouragement, and assistance throughout my time in the program, which have been invaluable to my academic progress and personal growth.

I gratefully acknowledge the financial support that made my doctoral studies and research possible. This work was supported in part by NIH grant R01GM148971, the Summer Research Seed Fund from the School of Data Science at UNC Charlotte, and the Faculty Research Grant from the Klein College of Science at UNC Charlotte. I am also grateful for the teaching assistantship support from the Department of Mathematics and Statistics and the Summer Fellowship from the Graduate School at UNC Charlotte.

Finally, I would like to thank my family and friends for their love, support, and constant presence throughout this journey. I am especially grateful to my parents

for their unconditional love, which has always been my greatest source of strength and comfort. I am also deeply grateful to my aunt, Xi Zhou, for her guidance, encouragement, and support. I would like to thank my friends Shanshan Wang and Su Xu for their companionship throughout my Ph.D. years. Last but not least, I am grateful to everyone who has brought kindness, warmth, and encouragement into my life along the way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Malaria

Malaria is a life-threatening infectious disease caused by protozoan parasites of the genus *Plasmodium*, which are transmitted to humans primarily through the bites of infected female *Anopheles* mosquitoes. Despite decades of control efforts, malaria remains a major global public health burden, particularly in tropical and subtropical regions. According to the World Health Organization, there were an estimated 282 million malaria cases and 610,000 malaria deaths worldwide in 2024 [2]. The burden of malaria is highly uneven across regions, with the WHO African Region accounting for the majority of global cases and deaths. As shown in Figure 1.1, malaria transmission is concentrated in sub-Saharan Africa, with additional burdens observed in parts of South America, South Asia, and Southeast Asia.

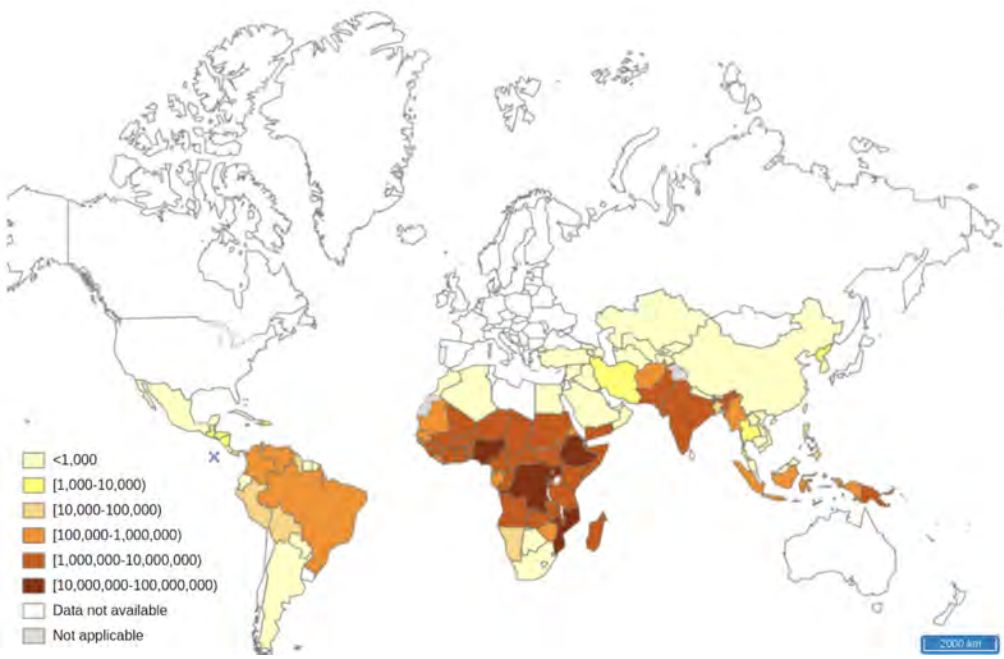


Figure 1.1: Estimated number of malaria cases by country (2024)

Source: World Health Organization

Several *Plasmodium* species can infect humans, including *P. falciparum*, *P. vi-*

vax, *P. malariae*, *P. ovale*, and *P. knowlesi* [3]. Among these species, *P. falciparum* is responsible for the most severe form of malaria and accounts for the majority of malaria-related deaths worldwide. *P. vivax* also contributes substantially to the global malaria burden and is notable for its ability to form dormant liver-stage parasites, which can reactivate and cause relapsing infections. These species differ in their geographic distribution, clinical severity, and biological characteristics, making malaria a complex disease from both clinical and biological perspectives.

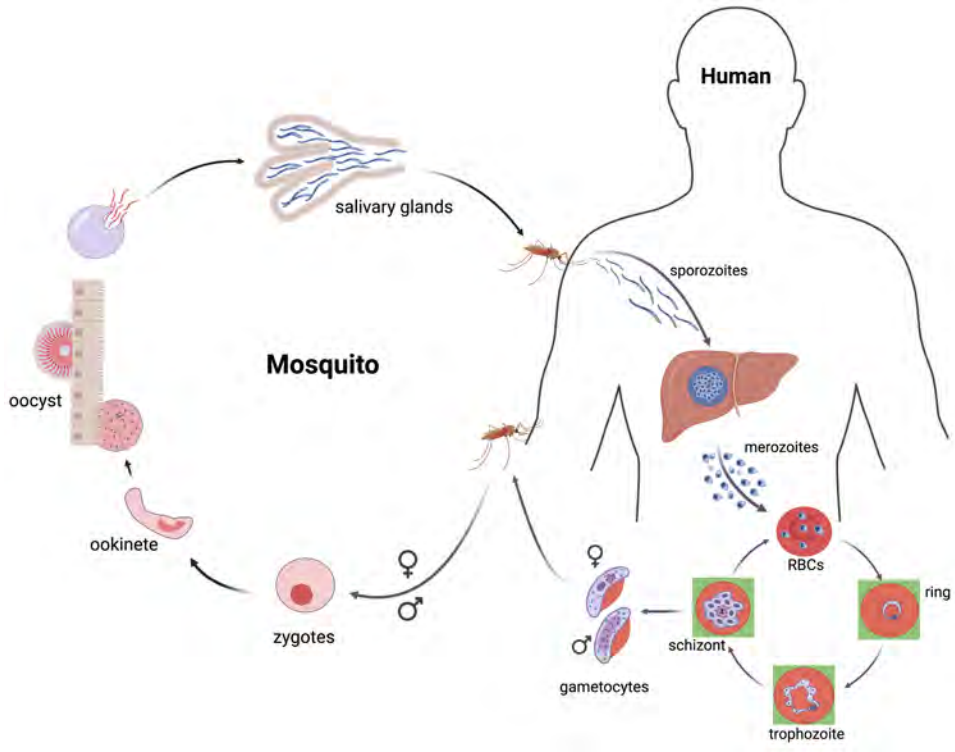


Figure 1.2: Life cycle of malaria parasites

As we can see in Figure 1.2, the *Plasmodium* life cycle is complex and involves both human and mosquito hosts. In humans, infection begins when sporozoites are injected into the bloodstream during a mosquito bite and migrate to the liver. After replication in hepatocytes, parasites are released into the bloodstream and invade red blood cells. During the intraerythrocytic developmental cycle, parasites progress through several morphologically and transcriptionally distinct stages, including the ring, trophozoite,

and schizont stages. This blood-stage cycle is responsible for the clinical symptoms of malaria and is therefore a major focus of molecular and transcriptomic studies.

1.2 RNA sequencing

RNA sequencing (RNA-seq) is a high-throughput sequencing technology used to measure transcript abundance at the genome-wide scale. Compared with earlier transcriptomic technologies, RNA-seq provides a digital, sequence-based measurement of RNA molecules and enables the quantification of gene expression, identification of novel transcripts, examination of alternative splicing, and comparison of transcriptional programs across biological conditions [4, 5]. In a typical RNA-seq experiment, RNA molecules are extracted from biological samples, converted into complementary DNA (cDNA), sequenced, and computationally processed to produce gene- or transcript-level count matrices. Because RNA-seq data are naturally represented as counts, downstream analyses often rely on statistical models that account for sequencing depth, biological variability, and overdispersion [6, 7]. In malaria research, RNA-seq and related transcriptomic technologies provide important tools for studying parasite development, host–parasite interactions, and stage-specific gene regulation during the *Plasmodium* life cycle.

1.2.1 Bulk RNA Sequencing

Bulk RNA sequencing measures the aggregate transcriptomic signal from a population of cells or organisms within a biological sample. In this approach, RNA is extracted from the entire sample, and sequencing reads are summarized to estimate the overall abundance of genes or transcripts. Bulk RNA-seq has been widely used for differential expression analysis, pathway analysis, and characterization of transcriptional changes across tissues, time points, disease states, or experimental perturbations [4, 5]. Its mature experimental protocols, relatively low cost per sample, and well-established statistical frameworks make it a powerful tool for studying gene

expression at the sample level.

A key limitation of bulk RNA-seq is that it averages expression across all cells present in the sample. When a sample contains multiple cell types, developmental stages, or cellular states, the observed bulk expression profile reflects a mixture of their individual transcriptional signals. Therefore, differences observed in bulk RNA-seq may result from true transcriptional regulation within a cell type, changes in cell-type or stage composition, or a combination of both. This limitation is particularly relevant in malaria studies, where blood-stage parasite populations may contain mixtures of ring, trophozoite, and schizont stages. As a result, bulk RNA-seq alone cannot directly resolve the stage-specific contributions underlying the observed expression profile, motivating the development of computational methods for estimating cellular or developmental composition from mixed transcriptomic data.

1.2.2 Single-cell RNA Sequencing

Single-cell RNA sequencing (scRNA-seq) extends transcriptomic profiling to the resolution of individual cells. Instead of measuring an averaged signal from a mixed population, scRNA-seq captures and sequences RNA molecules from individual cells, allowing researchers to characterize cellular heterogeneity, identify rare cell populations, reconstruct developmental trajectories, and study dynamic gene expression programs [8, 9]. Early scRNA-seq protocols demonstrated that whole-transcriptome expression profiling could be performed at the single-cell level [8]. Later advances, including droplet-based technologies, greatly increased the number of cells that could be profiled in parallel and made large-scale single-cell studies feasible [10, 11].

In malaria research, scRNA-seq has been used to profile individual *Plasmodium* parasites and to characterize transcriptional variation across parasite developmental stages and life-cycle transitions [12, 13]. These data are especially useful for studying the intraerythrocytic developmental cycle, during which parasites progress through transcriptionally distinct ring, trophozoite, and schizont stages. Single-cell transcrip-

tomic profiles can also serve as reference signatures for estimating stage composition in bulk RNA-seq samples. At the same time, scRNA-seq introduces additional statistical and computational challenges, including sparsity, technical noise, variable capture efficiency, batch effects, and the need for preprocessing, normalization, dimensionality reduction, clustering, and trajectory inference [14, 15, 16].

1.2.3 Comparison of Bulk and Single-cell RNA Sequencing

Bulk RNA-seq and single-cell RNA-seq provide complementary views of transcriptomic variation. Bulk RNA-seq is well suited for measuring sample-level transcriptional changes across many biological samples, whereas scRNA-seq is better suited for resolving heterogeneity among individual cells and identifying cell-type- or stage-specific expression patterns [4, 14, 17]. In this sense, bulk RNA-seq provides broad and stable measurements at the population level, while scRNA-seq provides higher-resolution information about the cellular composition and transcriptional states underlying those population-level signals.

The relationship between these two technologies is particularly important for complex biological systems. In mixed samples, the bulk expression profile can be viewed as a composite signal generated by multiple underlying cell types or developmental stages. Single-cell RNA-seq can help interpret this composite signal by providing reference profiles for individual cell populations or parasite stages. Computational deconvolution methods use these reference profiles to estimate the underlying cellular or developmental composition of bulk transcriptomic samples [18]. Thus, the integration of bulk and single-cell RNA-seq connects large-scale sample-level profiling with high-resolution cellular information.

This connection is central to the present dissertation. In malaria transcriptomic studies, bulk RNA-seq provides mixed expression profiles from parasite populations, while single-cell RNA-seq provides stage-specific reference information and enables the study of temporal gene expression dynamics at cellular resolution. Together, these

technologies motivate the two major methodological directions of this work: deconvolution of bulk transcriptomic samples using single-cell references, and functional modeling of single-cell temporal gene expression patterns.

1.3 Cell Composition Heterogeneity

Cell composition heterogeneity refers to variation in the relative abundance of different cell types, developmental stages, or cellular states across biological samples. In transcriptomic studies, this heterogeneity is especially important because different cell populations often have distinct gene expression profiles. Therefore, the gene expression measured from a heterogeneous sample reflects not only the transcriptional activity within each cell population, but also the relative proportions of those populations.

In bulk RNA-seq, the observed expression profile can be viewed as a mixture of cell-type- or stage-specific expression profiles. For a bulk sample containing K underlying cell populations, the expression of gene g can be approximately represented as

$$Y_g \approx \sum_{k=1}^{K} p_k X_{gk} + \epsilon_g \tag{1.1}$$

where Y_g denotes the observed bulk expression of gene g, X_{gk} denotes the expression of gene g in cell population k, p_k denotes the proportion of cell population k, and ϵ_g represents measurement noise and other unexplained variation. Under this framework, the bulk expression profile depends jointly on the cell-population-specific expression levels and the underlying cellular composition [19, 20].

Cell composition heterogeneity can complicate downstream analyses of bulk transcriptomic data. For example, a gene may appear to be differentially expressed between two conditions simply because the proportion of a cell type expressing that gene differs between the conditions. Conversely, true cell-type-specific transcriptional changes may be masked if they occur in a rare cell population or if changes in different

cell populations move in opposite directions. As a result, ignoring cell composition heterogeneity can lead to misleading conclusions in differential expression analysis, clustering, pathway analysis, and biomarker discovery[18, 20].

This issue is particularly relevant in malaria transcriptomic studies. During the intraerythrocytic developmental cycle, *Plasmodium* parasites progress through transcriptionally distinct ring, trophozoite, and schizont stages. Bulk RNA-seq samples from infected blood may contain mixtures of parasites at different developmental stages, especially when the parasite population is not perfectly synchronized. Therefore, observed differences in bulk parasite gene expression may reflect differences in stage composition, stage-specific gene regulation, or both. Estimating the developmental-stage composition of malaria samples is thus important for interpreting bulk transcriptomic profiles and for distinguishing biological regulation from compositional effects[21, 22].

Cell composition heterogeneity also motivates the integration of bulk and single-cell RNA-seq data. Single-cell RNA-seq provides high-resolution reference profiles for individual cell types or developmental stages, while bulk RNA-seq provides sample-level expression measurements across larger cohorts. By combining these two sources of information, computational methods can estimate the underlying cell-type or stage proportions in bulk samples and improve the biological interpretation of mixed transcriptomic signals [18, 20]. This provides the conceptual foundation for cell-type deconvolution methods, which are introduced in the following section.

1.4 Deconvolution Methods

Cell-type deconvolution methods aim to estimate the cellular or developmental composition of heterogeneous transcriptomic samples. In bulk RNA-seq studies, the observed expression profile is usually generated from a mixture of multiple cell types, developmental stages, or cellular states. Therefore, deconvolution methods seek to separate the mixed bulk signal into underlying cell-type-specific expression profiles

and their corresponding proportions.

A common formulation represents the bulk gene expression matrix as a linear mixture:

$$\mathbf{G} = \mathbf{CP} + \epsilon, \tag{1.2}$$

where $\mathbf{G}$ denotes the observed bulk expression matrix, with G_{ij} representing the expression level of gene i in bulk sample j. The matrix $\mathbf{C}$ denotes the cell-type- or stage-specific expression matrix, with $C_{i\ell}$ representing the expression level of gene i in cell type or stage ℓ. The matrix $\mathbf{P}$ denotes the cell-type or stage proportion matrix, with $P_{\ell j}$ representing the proportion of cell type or stage ℓ in bulk sample j. The term $\boldsymbol{\epsilon}$, with entries ϵ_{ij}, represents measurement noise and unexplained variation.

In many applications, the columns of $\mathbf{P}$ are constrained to be nonnegative and to sum to one, so that each column represents the composition of one bulk sample. This linear mixture model provides the basic mathematical framework for many transcriptomic deconvolution methods [18, 19, 20].

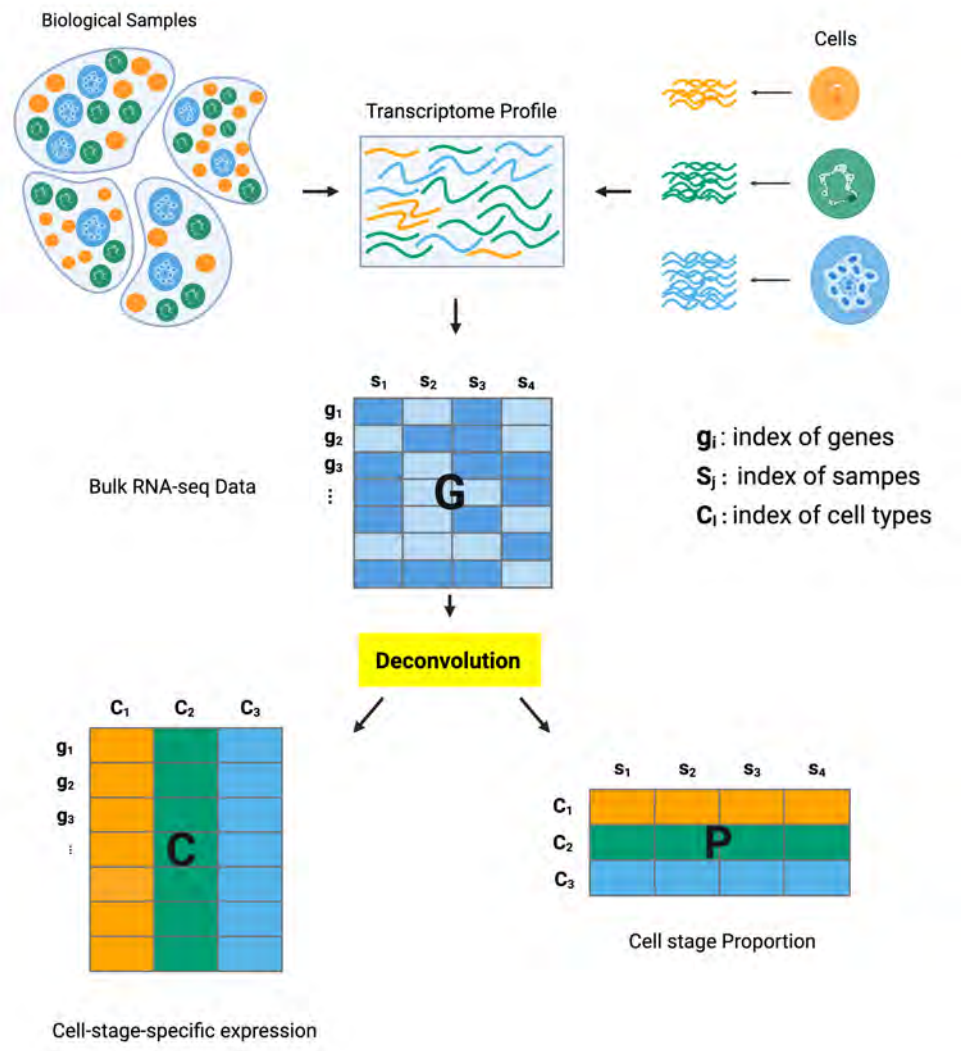


Figure 1.3: Illustration of transcriptomic deconvolution. The observed bulk RNA-seq expression matrix (G) is modeled as a mixture of cell-type- or stage-specific expression profiles (C) and sample-specific cell-type or stage proportions (P), with residual noise represented by (ϵ).

As shown in Figure 1.3, deconvolution links the observed bulk transcriptomic profile to two biologically interpretable components: the expression signatures of the underlying cell populations and their relative abundances across samples. Depending on whether external reference profiles are available, deconvolution methods can be broadly divided into reference-based and reference-free approaches.

Reference-based methods assume that cell-type-specific expression profiles are available from external reference data, such as purified cell populations or single-cell

RNA-seq. Under this setting, (C) is treated as known or estimated from the reference, and the main goal is to estimate the proportion matrix (P). Methods such as CIBERSORT, CIBERSORTx, MuSiC, Scaden, and BayesPrism belong to this category [19, 20, 23, 24, 25]. These methods differ in their statistical assumptions and computational strategies. For example, CIBERSORT uses support vector regression to estimate cell fractions from bulk expression profiles, MuSiC uses multi-subject single-cell references to estimate cell-type proportions, Scaden applies deep learning to predict cellular composition, and BayesPrism uses a Bayesian framework to integrate bulk and single-cell RNA-seq data.

Reference-based methods are powerful when a high-quality and biologically matched reference is available. Single-cell RNA-seq has made this approach increasingly useful because it provides high-resolution cell-type- or stage-specific expression profiles. However, reference-based methods can be sensitive to differences between the bulk data and the reference data, including batch effects, platform differences, missing cell types, cell-state differences, and biological mismatch between the reference and target samples.

Reference-free methods do not require external cell-type-specific reference profiles. Instead, they attempt to estimate both the expression matrix (C) and the proportion matrix (P) directly from the bulk expression matrix (G). This setting is also called complete deconvolution because both unknown components of the mixture model are estimated. Reference-free methods are useful when suitable reference profiles are unavailable or unreliable, but they are statistically more challenging because the factorization problem is generally ill-posed and may not have a unique solution. Methods such as CDSeq, STdeconvolve, and GSNMF address this problem by imposing additional modeling assumptions, latent structure, or geometric constraints to improve identifiability and interpretability [26, 27, 1].

In malaria transcriptomic studies, deconvolution is particularly relevant because

bulk RNA-seq samples may contain mixtures of parasites at different developmental stages. In this context, the columns of (C) can be interpreted as stage-specific expression profiles for ring, trophozoite, and schizont stages, while the columns of (P) represent the corresponding stage proportions in each bulk sample. Accurate estimation of these proportions can help distinguish changes in parasite stage composition from true stage-specific transcriptional regulation. Therefore, deconvolution provides a key computational tool for interpreting mixed malaria transcriptomic data.

Among the deconvolution approaches described above, reference-free methods are especially relevant when reliable cell-type- or stage-specific reference profiles are unavailable or when the available references may be affected by batch, platform, or biological differences. The following section introduces geometric structure-guided nonnegative matrix factorization, a reference-free complete deconvolution framework that builds on the linear mixture model in Equation 1.2 and incorporates geometric information from the bulk expression data to improve identifiability and interpretability.

1.5 Geometric Structure Guided Nonnegative Matrix Factorization

Mathematically, reference-free deconvolution can be solved as a nonnegative matrix factorization (NMF) problem:

$$(\mathbf{C}^*, \mathbf{P}^*) = \arg \min_{\mathbf{C}\geq 0, \mathbf{P}\geq 0} \frac{1}{2}\|\mathbf{G} - \mathbf{CP}\|_F^2 + \mathbf{1}_T(\mathbf{P}) \tag{1.3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $\mathbf{1}_T(\mathbf{P})$ is an indicator function that enforces constraints on the proportion matrix, such as nonnegativity and column-wise sum-to-one constraints.

However, the NMF is strongly ill-posed and solutions are generally not unique: if

$(\mathbf{C}^*, \mathbf{P}^*)$ is a local minimum , then for any $\mathbf{\Omega} \in \mathbb{R}^{k \times k}$,

$$\hat{\mathbf{C}} = \mathbf{C}^*\mathbf{\Omega} \quad \text{and} \quad \hat{\mathbf{P}} = \mathbf{\Omega}^{-1}\mathbf{P}^* \tag{1.4}$$

are also solutions, as long as their non-negativity is satisfied.

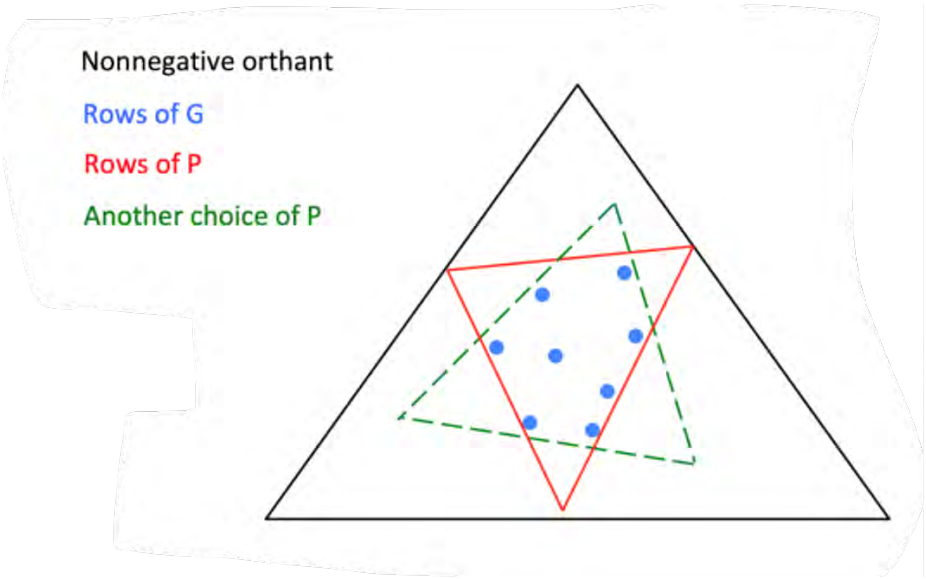


Figure 1.4: Convex-hull view of NMF non-uniqueness
Reproduced from Chen et al.[1]

The non-uniqueness of the basic NMF formulation can also be understood from a geometric perspective. As shown in Figure 1.4, the blue dots represent the rows of $\mathbf{G}$, and the black triangle represents the nonnegative orthant. The NMF problem can then be interpreted as finding a simplex, whose vertices correspond to the rows of $\mathbf{P}$, that contains all observed rows of $\mathbf{G}$ within the nonnegative orthant. However, without additional constraints, such a simplex is not unique. For example, both the red solid triangle and the green dashed triangle are feasible choices of $\mathbf{P}$ because they both enclose the observed data points. Therefore, multiple factorizations can explain the same data matrix $\mathbf{G}$, leading to non-identifiable and biologically ambiguous estimates of cell-type proportions and cell-type-specific expression profiles.

The non-uniqueness of NMF solutions can substantially affect downstream sta-

tistical analyses and lead to ambiguous biological interpretations. Therefore, it is necessary to restrict the feasible search space of the factor matrices in order to improve solution identifiability and interpretability. The uniqueness of an NMF solution is formally defined as follows [28].

Definition 1 (Uniqueness of NMF solution) The solution $(\mathbf{C}^*, \mathbf{P}^*)$ of NMF is unique, or identifiable, if and only if for any other solution $(\bar{\mathbf{C}}, \bar{\mathbf{P}})$, there exists a permutation matrix $\mathbf{\Pi} \in \{0,1\}^{k\times k}$ and a diagonal scaling matrix $\mathbf{S}$ with positive diagonal matrix such that

$$\bar{\mathbf{C}} = \mathbf{C}^*\mathbf{\Pi}\mathbf{S} \quad \text{and} \quad \bar{\mathbf{P}} = \mathbf{S}^{-1}\mathbf{\Pi}^\intercal\mathbf{P}^*.$$

Further results in [29, 28] indicate that, under certain conditions [30, 31, 32, 33], the NMF solution can indeed be unique. Two notable results among them are:

Theorem 1 (Strong identifiability condition) Assuming $k = \text{rank}(\mathbf{G})$, $\epsilon = 0$, if problem admits a solution, for which both $\mathbf{C}^\intercal$ and $\mathbf{P}$ are separable matrices, then the solution is unique.

Theorem 2 (Weak identifiability condition) Assuming $k = \text{rank}(\mathbf{G})$, $\epsilon = 0$, if both $\mathbf{C}^\intercal$ and $\mathbf{P}$ are sufficiently scattered, then problem admits a unique solution.

The key question is how these theoretical identifiability conditions can be connected to bulk RNA-seq deconvolution. In gene expression data, this connection can be established through the concept of marker genes [1]. A marker gene for a given cell stage is expected to be dominantly expressed in that stage and rarely expressed in other stages. Therefore, if a gene is a marker for stage r, its expression pattern across bulk samples should be highly correlated with the proportion of that stage. In the ideal noiseless case, this means that the corresponding row of $\mathbf{C}$ is dominated by the r-th component, and the row of $\mathbf{G}$ is proportional to the r-th row of $\mathbf{P}$.

This observation provides a biological interpretation of the geometric structure in

NMF. Since $\mathbf{G} = \mathbf{CP}$, each row of $\mathbf{G}$ can be represented as a nonnegative combination of the rows of $\mathbf{P}$, where the corresponding row of $\mathbf{C}$ provides the coefficients. Thus, rows of $\mathbf{G}$ associated with marker genes from the same cell stage tend to cluster around the corresponding direction of $\mathbf{P}$. This motivates the use of marker genes to identify the geometric directions associated with different cell stages. Moreover, because rows of $\mathbf{C}$ can be interpreted as coefficient vectors of rows of $\mathbf{G}$ under the basis given by $\mathbf{P}$, the geometric structure observed in $\mathbf{G}$ can be transferred to $\mathbf{C}$ through additional constraints. This idea forms the basis for using marker-gene information to improve the identifiability and interpretability of GS-NMF.

Based on this observation, GS-NMF follows a structure-exploring and structure-preserving strategy. First, the rows of $\mathbf{G}$ are clustered according to their expression patterns across bulk samples, so that potential marker genes for each cell stage can be identified. Then, the geometric structure identified from $\mathbf{G}$ is imposed on the coefficient matrix $\mathbf{C}$ through two regularization terms: a solvability constraint and a manifold constraint.

The solvability constraint uses the selected marker genes to improve identifiability. For a marker gene associated with stage r, the corresponding row of $\mathbf{C}$ is expected to be close to the direction of the r-th unit vector $\mathbf{e}_r^\intercal$, because this gene should be dominantly expressed in stage r and rarely expressed in other stages. Therefore, this constraint encourages the marker-gene rows of $\mathbf{C}$ to be sufficiently scattered toward different cell-stage directions, which helps reduce the ambiguity of the NMF solution. Then the solvability penalty is defined as

$$\mathcal{F}_1(\mathbf{C}) = \frac{\lambda_1}{2} \sum_{r=1}^{k} \sum_{i \in S_r} d_{\text{eisen}}(\mathbf{C}(i), \mathbf{e}_r^\intercal)^2 \tag{1.5}$$

where λ_1 controls the strength of the solvability constraint. This term encourages the marker-gene rows of $\mathbf{C}$ to be close to the corresponding coordinate directions, thereby

improving identifiability.

The manifold constraint preserves the correlation structure observed in the rows of $\mathbf{G}$. If two genes have similar expression patterns across bulk samples, their coefficient representations in $\mathbf{C}$ should also be similar. Let ω_{ij} denote the edge weight between genes i and j in the adjacency matrix constructed from $\mathbf{G}$. The manifold penalty is defined as

$$\mathcal{F}_2(\mathbf{C}) = \frac{\lambda_2}{2} \sum_{j=1}^{N} \sum_{i=1}^{N} \omega_{ij} d_{\text{eisen}}(\mathbf{C}(i), \mathbf{C}(j))^2 \tag{1.6}$$

where λ_2 controls the strength of the manifold constraint. This term encourages rows of $\mathbf{C}$ corresponding to highly correlated rows of $\mathbf{G}$ to remain close under the same correlation-based distance. Here, $d_{\text{eisen}}(\cdot, \cdot)$ denotes the Eisen cosine correlation distance.

Combining the solvability condition and manifold constraint, we propose the following geometric structure guided nonnegative matrix factorization (GSNMF) model. For illustration convenience, we define the set $T := \{\mathbf{Z} \in \mathbb{R}_+^{k \times n}, \mathbf{1}^\intercal \mathbf{Z} = \mathbf{1}^\intercal\}$ and the indicator function $\mathbf{1}_T$ as $\mathbf{1}_T(\mathbf{Z}) = 0$ if $\mathbf{Z} \in T$ while $\mathbf{1}_T(\mathbf{Z}) = \infty$ otherwise. With these notations, solving for $\mathbf{C}$ and $\mathbf{P}$ becomes the optimization problem:

$$\min_{\mathbf{C} \geq 0, \mathbf{P} \geq 0} \frac{1}{2} \|\mathbf{G} - \mathbf{C}\mathbf{P}\|_F^2 + \mathcal{F}(\mathbf{C}) + \mathbf{1}_T(\mathbf{P}) \tag{1.7}$$

where $\mathcal{F}(\mathbf{C}) = \mathcal{F}_1(\mathbf{C}) + \mathcal{F}_2(\mathbf{C})$, and the third term $\mathbf{1}_T(\mathbf{P})$ simply means sum-to-one conditions on columns of $\mathbf{P}$, or column stochasticity, since the sum of cellular proportions in each tissue sample is supposed to be one.

The GS-NMF framework addresses the deconvolution of bulk RNA-seq data by estimating cell-stage proportions and stage-specific expression profiles. However, bulk

deconvolution does not directly describe how gene expression changes over developmental time. For malaria parasites, the intraerythrocytic developmental cycle is a continuous temporal process, and single-cell RNA-seq data provide an opportunity to study gene expression dynamics at a finer resolution. Therefore, we next shift from bulk-level deconvolution to single-cell temporal modeling, focusing on how genes vary along pseudotime.

1.6 Single-cell Temporal Gene Expression Dynamics

Many biological processes are inherently dynamic. During development, differentiation, infection, and disease progression, cells gradually transition from one state to another while activating or repressing specific transcriptional programs. As we methioned in Section 1.2, Single-cell RNA sequencing(scRNA-seq) provides an opportunity to study these transitions at high resolution by measuring gene expression profiles from individual cells [34, 14]. Compared with bulk RNA-seq, which measures average expression across a population of cells, scRNA-seq preserves cell-to-cell variability and allows researchers to characterize heterogeneous and continuous biological states [14].

Temporal gene expression dynamics are especially important in systems where biological states change gradually rather than abruptly. A gene may be highly expressed at an early stage, gradually increase or decrease over time, peak at an intermediate stage, or show transient expression during a narrow developmental window. These diverse temporal patterns can reflect different biological roles, such as initiating a transition, maintaining a cell state, or regulating stage-specific functions. Therefore, identifying and characterizing temporal expression patterns is a key step in understanding dynamic biological systems [34, 35].

A central goal of single-cell temporal analysis is to identify genes whose expression changes systematically along an underlying biological trajectory. These temporally dynamic genes often reflect key regulatory programs, stage-specific biological func-

tions, and coordinated transcriptional modules. By studying the temporal behavior of genes, researchers can gain insight into how transcriptional programs are organized and how cells progress through different biological states [35].

In the context of malaria, the parasite life cycle involves coordinated transcriptional changes across developmental stages. Genes expressed during the ring, trophozoite, and schizont stages may participate in distinct biological processes related to parasite growth, metabolism, replication, and host-cell interaction [13, 36]. Single-cell malaria datasets therefore provide a useful setting for studying temporal gene expression dynamics and discovering gene modules associated with parasite development [13].

Despite its potential, modeling temporal gene expression from scRNA-seq data remains challenging. Single-cell expression profiles are high-dimensional, noisy, sparse, and often contain a large fraction of zero counts. Gene expression counts may also exhibit overdispersion, where the observed variability exceeds what is expected under a simple Poisson model [37, 38]. These features make it difficult to directly model temporal expression patterns or cluster genes based on raw expression profiles. Therefore, effective temporal analysis requires statistical methods that can extract informative temporal features, reduce noise, and summarize gene expression dynamics in a biologically meaningful way.

1.7 Pseudotime Estimation

In many single-cell studies, the true biological time of each cell is unknown. Although cells may be collected from different experimental time points, individual cells within the same sample can still be at different biological stages because of asynchronous development, heterogeneous responses, or technical variation. As a result, experimental time alone may not accurately represent the underlying progression of a dynamic biological process. Pseudotime estimation was developed to address this problem by ordering cells along an inferred trajectory based on their transcriptomic similarity [34, 39].

Pseudotime is a latent variable that represents the relative position of each cell along a biological trajectory. Cells assigned smaller pseudotime values are interpreted as being closer to the beginning of the process, while cells assigned larger pseudotime values are considered to be at later stages. Unlike real chronological time, pseudotime does not necessarily correspond to physical time. Instead, it provides an inferred ordering that captures gradual transcriptional changes across cells. This ordering allows researchers to study continuous gene expression dynamics even when cells are sampled only at a limited number of time points or when the biological process is not synchronized.

A variety of computational methods have been proposed for trajectory inference and pseudotime estimation. Early approaches, such as Monocle, order cells along trajectories constructed from dimension-reduced expression data and have been widely used to study cell differentiation and developmental processes [34]. Diffusion pseudotime uses diffusion maps to capture gradual transitions and is particularly useful for reconstructing branching lineage structures [40]. Slingshot combines clustering with simultaneous principal curves to infer lineage-specific trajectories and pseudotime values in reduced-dimensional space [41]. Other graph-based methods, such as PAGA, represent the connectivity among cell populations and provide a topology-preserving framework for studying complex developmental processes [42].

Although pseudotime estimation provides an important foundation for temporal single-cell analysis, it also introduces statistical and computational challenges. Different trajectory inference methods may produce different cell orderings depending on preprocessing choices, dimensionality reduction methods, clustering results, and assumptions about trajectory structure. In addition, pseudotime is an estimated quantity and may contain uncertainty, especially in noisy or sparsely sampled regions of the trajectory. Therefore, downstream analysis should account for the fact that pseudotime is not directly observed but inferred from high-dimensional gene expres-

sion data.

1.8 Functional Modeling of Temperal Gene Expression

After pseudotime has been estimated, gene expression can be modeled as a function of the inferred temporal ordering. For each gene, the objective is to characterize how its expression level changes as cells progress along a biological trajectory. From this perspective, temporal gene expression can be viewed as a functional object, where the observed expression values across cells are noisy measurements of an underlying expression pattern. Functional modeling therefore provides a natural framework for studying dynamic transcriptional changes in single-cell data.

Temporal gene expression patterns are often highly nonlinear. Some genes may show monotonic increases or decreases along pseudotime, while others may be transiently activated during a specific developmental window or peak at an intermediate stage. These diverse patterns cannot always be captured by simple linear models. Therefore, flexible modeling strategies are needed to represent complex expression dynamics and identify genes whose expression changes significantly along pseudotime.

Several statistical approaches have been developed for modeling gene expression as a smooth function of time or pseudotime. Spline-based models and generalized additive models (GAMs) are commonly used because they provide flexible representations of nonlinear temporal trends. Early pseudotime-based methods, such as Monocle, model gene expression as a smooth nonlinear function of pseudotime using generalized additive models implemented through the VGAM framework [34]. Building on this idea, tradeSeq uses negative binomial generalized additive models (NB-GAMs) to model gene expression along one or multiple lineages and provides statistical tests for trajectory-associated differential expression [35]. PseudotimeDE further considers the uncertainty introduced by pseudotime estimation by using subsampling-based procedures and fitting NB-GAM or zero-inflated negative binomial GAM (ZINB-GAM)

models to obtain well-calibrated p-values for detecting genes associated with pseudotime [43]. Other methods have also been proposed for temporal or pseudotime-based expression analysis. For example, NBAMSeq uses a negative binomial additive model to analyze time-course RNA-seq data [44], while ImpulseDE2 uses an impulse model to capture transient activation and repression patterns in temporal sequencing data [45]. Together, these methods demonstrate the importance of flexible functional modeling for characterizing nonlinear gene expression dynamics along biological trajectories.

From the perspective of functional data analysis, temporal expression curves can be represented using basis functions. Common choices include polynomial bases, B-splines, Fourier bases, and other smooth basis expansions [46, 47]. A basis representation expresses the temporal expression pattern of each gene as a weighted combination of predefined functions. The estimated weights or coefficients summarize the shape of the gene's expression trajectory and can be used as low-dimensional features for downstream analysis.

This representation is especially useful for single-cell temporal gene expression analysis because raw expression profiles are often sparse, noisy, and high-dimensional. Instead of directly analyzing raw count profiles, functional representations summarize the underlying temporal trend of each gene in a more interpretable form. These fitted temporal functions can be used to identify genes with dynamic expression patterns, compare expression trajectories across genes, and provide informative features for downstream clustering and gene module discovery. However, modeling single-cell temporal expression remains statistically challenging because the observed data are discrete counts and often exhibit overdispersion, excess zeros, and uncertainty from pseudotime estimation. Therefore, effective temporal modeling methods should be flexible enough to capture nonlinear expression patterns while also accounting for the distributional characteristics of scRNA-seq data.

1.9 Functional Clustering for Gene Module Discovery

After temporal gene expression patterns have been modeled, a natural next step is to group genes with similar dynamic behaviors into gene modules. In many biological processes, genes do not act independently; instead, groups of genes are often co-regulated and participate in related biological pathways or stage-specific functions. Identifying such gene modules can provide a higher-level interpretation of temporal transcriptomic data and reveal coordinated regulatory programs underlying biological transitions.

Traditional clustering methods, such as hierarchical clustering, k-means clustering, and model-based clustering, are commonly used to group genes based on expression similarity. However, when applied directly to raw single-cell expression profiles, these methods may be sensitive to noise, sparsity, and cell-to-cell variability. This issue is especially important in temporal single-cell analysis, where the main interest is not only the expression level at individual cells, but also the overall shape of the expression trajectory along pseudotime. Therefore, clustering genes based on their underlying temporal patterns can be more informative than clustering genes based only on raw expression values.

Functional clustering provides a framework for grouping genes according to the shapes of their temporal expression functions. Instead of treating each gene as a vector of raw expression measurements, functional clustering represents each gene by a fitted curve or a set of functional features. Genes with similar temporal trends, such as early activation, late activation, transient expression, or monotonic changes, can then be grouped into the same cluster. This approach reduces the influence of noisy observations and focuses on the major patterns of temporal variation. Functional clustering has been widely used in time-course gene expression studies to identify groups of genes with shared dynamic profiles [48, 49, 50].

Several methods have been developed for clustering or summarizing temporal gene

expression patterns. Fuzzy c-means clustering has been widely used for time-course gene expression analysis because it allows genes to have partial membership in multiple clusters rather than forcing each gene into a single module. For example, Mfuzz applies fuzzy (c)-means clustering to group genes into expression modules, which is useful when genes may participate in multiple biological processes or show overlapping temporal patterns [49, 51]. STEM was designed for short time-series gene expression data and clusters genes according to predefined temporal expression profiles [50]. Functional principal component analysis (FPCA) provides another approach by representing temporal expression curves through a small number of principal functional components; the resulting FPCA scores can then be used as low-dimensional features for clustering genes with similar temporal trajectories [46, 47]. Other approaches combine smoothing, regression, or functional data analysis with clustering to group genes according to estimated temporal trajectories rather than raw measurements [48, 52]. In single-cell trajectory analysis, methods such as tradeSeq first fit smooth gene expression functions along pseudotime using negative binomial generalized additive models; although tradeSeq is primarily designed for trajectory-based differential expression analysis, the fitted temporal patterns or predicted expression values can also support downstream clustering of dynamic genes into temporal modules [35]. Together, these methods illustrate the importance of using temporal structure and functional representations when identifying gene modules.

Overall, functional clustering provides a bridge between statistical modeling and biological interpretation. By grouping genes according to their modeled temporal expression patterns, it allows researchers to move beyond individual gene-level results and identify coordinated transcriptional programs. These gene modules can then be further analyzed through functional annotation, pathway enrichment, or comparison with known stage-specific markers to better understand the biological processes underlying temporal single-cell data.

1.10 Overview of the Dissertation

The remainder of this dissertation is organized as follows.

Chapter 2 presents GSNMF+, a geometry-guided nonnegative matrix factorization framework with data augmentation for robust deconvolution of bulk RNA-seq data. Building upon the original GS-NMF framework, GSNMF+ introduces several key improvements, including a more universal framework extension, an annotation strategy for estimated components, and a gradient-descent-based optimization procedure. These improvements allow the method to be more flexibly applied to malaria bulk RNA-seq data and improve the interpretability of the estimated cell-stage proportions. This chapter first introduces the biological and statistical motivation for complete deconvolution of malaria bulk transcriptomic data. It then describes the framework design, including the construction of augmented reference mixtures, marker-gene-based geometric constraints, component annotation, and the optimization strategy used to estimate cell-stage proportions and stage-specific expression profiles. The chapter also summarizes the datasets used in the simulation and real-data studies, followed by simulation results, stability analyses, and applications to real bulk RNA-seq data. Finally, the chapter discusses the overall performance of GSNMF+ and its advantages for improving robustness and interpretability in bulk RNA-seq deconvolution.

Chapter 3 introduces BetaDE, a statistical framework for modeling single-cell temporal gene expression dynamics. While Chapter 2 focuses on deconvolving bulk RNA-seq samples into discrete malaria cell-stage proportions, Chapter 3 shifts the focus to single-cell RNA-seq data, where gene expression can be studied continuously along pseudotime. This chapter describes the construction of beta kernel basis functions, the statistical models used for gene expression counts, model selection procedures, differential expression detection, and functional clustering of significant genes. The proposed method is evaluated through simulation studies and validation using known time-point data, and is further applied to real single-cell malaria datasets. The chap-

ter concludes by discussing how temporal single-cell modeling complements bulk deconvolution by providing a finer-resolution view of gene expression changes across the Malaria intraerythrocytic developmental cycle.

Chapter 4 outlines several directions for future research. Building on the methodological developments in Chapters 2 and 3, this chapter discusses potential extensions of GSNMF+ and BetaDE, including improvements in model flexibility, incorporation of additional biological information, and broader applications to other transcriptomic datasets. It also summarizes the main contributions of this dissertation and highlights remaining challenges in bulk RNA-seq deconvolution and single-cell temporal gene expression modeling.

CHAPTER 2: GSNMF+: A Geometry-Guided NMF Framework with Data Augmentation for Robust Deconvolution of Bulk RNAseq Data

2.1 Introduction

Most biological tissue samples, including tumors, brain tissues, blood samples, and infected host tissues, are heterogeneous mixtures of multiple cell populations whose relative abundances vary across individuals, experimental conditions, disease states, and developmental time points. Although bulk RNA-seq provides an efficient way to measure transcriptomic signals from such samples, the observed gene expression profile represents an averaged signal across all constituent cell types or cell states. Therefore, changes observed in bulk gene expression may reflect shifts in cellular composition, true transcriptional regulation within specific cell populations, or a combination of both. This mixture structure complicates downstream biological interpretation, especially in differential expression analysis, where the goal is often to distinguish cell-intrinsic transcriptional changes from changes driven by cell-type or cell-stage abundance.

Computational deconvolution of bulk transcriptomic data has therefore emerged as an important strategy for estimating cell-type proportions and cell-type-specific gene expression profiles from mixed samples. Deconvolution methods are broadly categorized into reference-based and reference-free approaches. Reference-based methods, including CIBERSORT [19], CIBERSORTx [23], MuSiC [20], and BayesPrism [25], use cell-level or purified cell-type reference profiles to estimate cellular proportions in bulk mixtures. These methods can be effective when high-quality and biologically matched reference data are available. However, their performance depends heavily on the quality, completeness, and relevance of the reference panel. Reference-free methods, such as CDSeq [26], CellDistinguisher [53], LinSeed [54], and TOAST [55], instead attempt to estimate both cellular proportions and cell-type-specific expres-

sion profiles directly from bulk data. These approaches are particularly useful when suitable reference data are unavailable, incomplete, or unreliable.

A compelling application of bulk RNA-seq deconvolution arises in malaria transcriptomic studies. Despite the availability of preventive measures and treatment strategies, malaria continues to impose a severe global health burden, particularly in low-income settings. More than 200 *Plasmodium* species have been formally described, among which five species are responsible for malaria in humans on a global scale: *P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale*, and *P. knowlesi* [3]. Malaria research has historically focused on *P. falciparum* because of its high mortality and high prevalence in sub-Saharan Africa. However, recent studies have shown that *P. vivax* can also cause severe and potentially fatal disease, highlighting the need for expanded research attention to non-*falciparum* malaria [56, 57].

Within a single patient blood sample, bulk RNA-seq captures a mixture of co-occurring intraerythrocytic parasite stages. As a result, observed changes in bulk gene expression may arise from true transcriptional regulation within a specific developmental stage or from changes in the relative abundance of different parasite stages. For example, in chloroquine-treated *P. vivax* bulk RNA-seq studies, elevated expression of stress-response genes may reflect direct transcriptional activation within trophozoites. Alternatively, it may result from preferential clearance of trophozoites and subsequent enrichment of other stages that constitutively express these genes at higher levels. Distinguishing between these scenarios is essential for accurate biological interpretation, particularly in the context of drug response and developmental regulation.

A further complication arises from the limited availability of species-matched reference data. When applying reference-based deconvolution to bulk RNA-seq data from *P. vivax* infections, the most readily available single-cell RNA-seq reference profiles are often derived from *P. falciparum*. This introduces biologically meaningful

interspecies discrepancies that can lead to substantial inaccuracies in deconvolution. Moreover, even within the same species, single-cell RNA-seq datasets generated under different experimental conditions may show considerable variability, suggesting that no single reference dataset is universally representative.

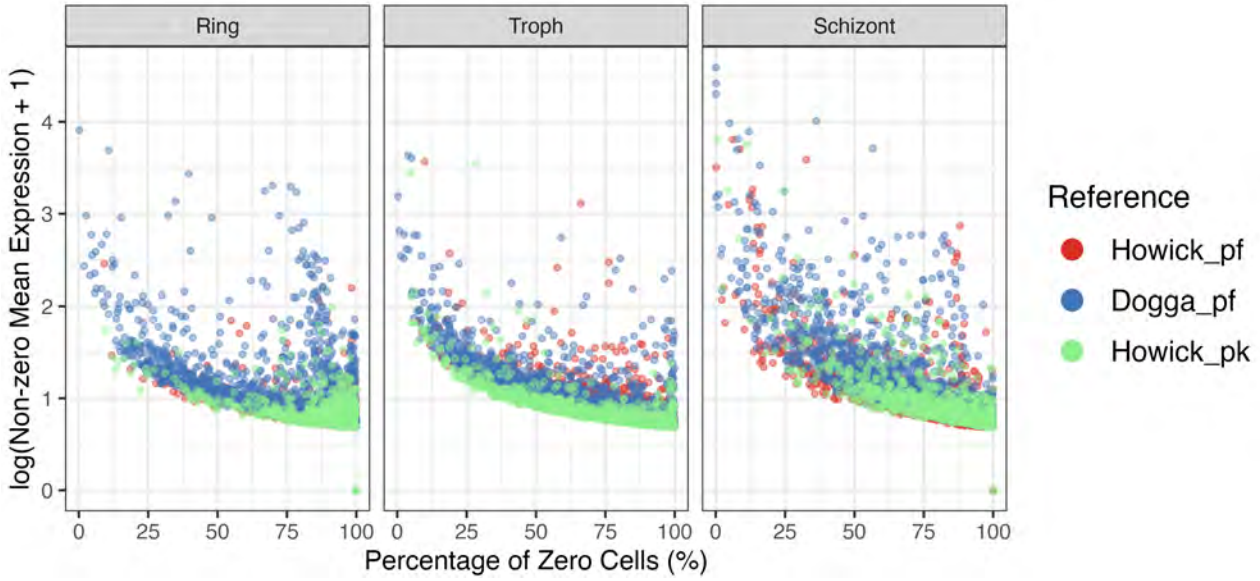


Figure 2.1: Reference heterogeneity across single-cell datasets. Comparison of percentage of zero expressions and log-transformed nonzero mean expression across three single-cell reference datasets. Each panel corresponds to one parasite stage, and each point represents one gene. Differences among references indicate heterogeneity in sparsity and expression magnitude across these single-cell datasets.

This issue is illustrated by comparing the single-cell reference datasets used in this study. Figure 2.1 shows the relationship between the percentage of zero expression of a gene across cells and the log-transformed nonzero mean expression at ring, trophozoite, and schizont stages, respectively, across three references [13, 58]. The references differ in both sparsity and expression magnitude, with some datasets showing lower nonzero mean expression and higher zero proportions for many genes. Such heterogeneity can potentially affect reference-based deconvolution and motivates the development of methods that are robust to reference choice and sparse expression patterns.

Nonnegative matrix factorization (NMF) provides a natural mathematical frame-

work for complete deconvolution because both gene expression levels and cellular or stage proportions are nonnegative. In this setting, the observed bulk expression matrix is approximated by the product of a cell-type-specific or stage-specific expression matrix and a proportion matrix. However, standard NMF is highly ill-posed and generally lacks solution uniqueness. Without additional structural constraints, the factorization can be non-identifiable and unstable, leading to biologically ambiguous estimates.

To address these challenges, geometry-guided NMF incorporates biological and geometric structure into the factorization framework. The original GS-NMF method uses marker-gene information to impose a solvability constraint and uses a manifold regularization term to preserve the local correlation structure of the expression space [1]. Although GS-NMF improves identifiability and interpretability, its performance may degrade when observed bulk samples do not adequately span the underlying proportion space. In addition, its reliance on the alternating direction method of multipliers can introduce implementation complexity.

In this chapter, we introduce GSNMF+, an enhanced and generalized extension of GS-NMF [1] for robust bulk RNA-seq deconvolution. GSNMF+ introduces several key improvements, which we will discuss details in Section 2.2. Section 2.3 introduces the single-cell and bulk RNA-seq datasets used in this study. Section 2.4 presents simulation results, followed by stability analyses in Section 2.5. Section 2.6 applies the proposed method to real bulk RNA-seq data. Finally, Section 2.7 summarizes the main findings and discusses the advantages and limitations of GSNMF+.

2.2 Study Design

GSNMF+ is developed as an extension of the original GS-NMF framework to improve the robustness, interpretability, and computational implementation of bulk RNA-seq deconvolution. The original GS-NMF framework is theoretically effective when the observed bulk samples contain sufficiently diverse cellular or stage com-

positions, so that the geometry of the mixture space is well represented. However, in real applications, the observed bulk samples may be highly unbalanced or nearly singular in their cellular composition. For example, most samples may be dominated by one stage, such as ring, trophozoite, or schizont, and therefore may not adequately span the full proportion space. In such cases, the weak identifiability condition may not be well supported by the observed data alone, leading to unstable or biologically ambiguous deconvolution results.

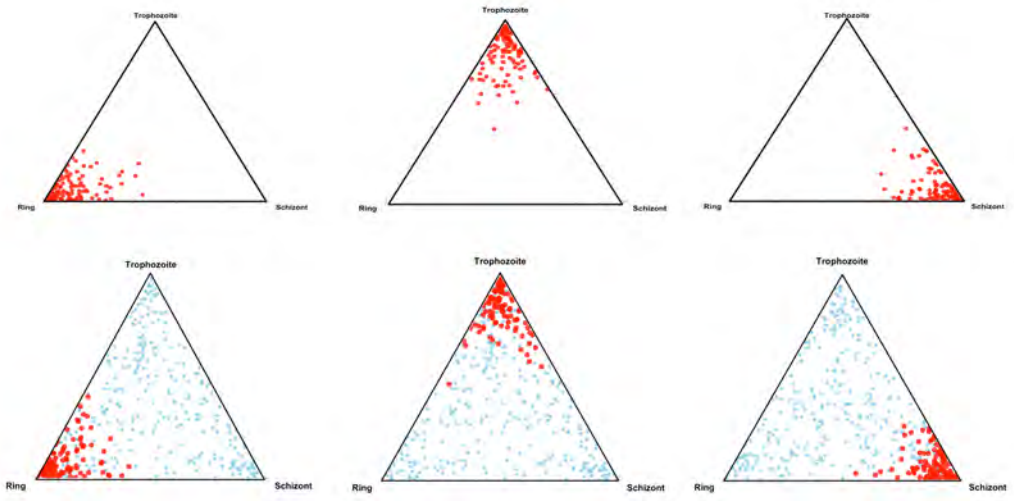


Figure 2.2: Key improvements of GSNMF+. Red points represent observed bulk mixtures whose stage compositions may be concentrated near one vertex of the simplex, corresponding to ring-, trophozoite-, or schizont-dominant samples. Blue points represent artificially generated augmented mixtures. Combining observed and augmented samples improves coverage of the proportion space and helps stabilize the geometry-guided NMF decomposition.

To address these limitations, GSNMF+ introduces three major improvements: a universal framework extension based on data augmentation, an estimated component annotation strategy, and a multiplicative-update-based optimization procedure. Figure 2.2 provides a schematic illustration of the augmentation idea. Each point represents one bulk mixture in the stage-proportion simplex. The red points represent observed mixtures that are concentrated near a single stage direction, while the blue points represent artificially generated augmented mixtures. By combining observed

and augmented mixtures, the data are better distributed across the simplex and can more effectively support the geometric constraints used in GS-NMF.

2.2.1 Universal Framework Extension Using Data Augmentation

Identifiability of the NMF factorization requires that the sample compositions provide sufficient geometric information. Informally, the columns of the proportion matrix $\mathbf{P}$ should be well scattered across the composition simplex rather than confined to a small region. In the strongest separable case, each cell stage should be close to dominant in at least one sample. Under such scatter conditions, the nonnegative factorization becomes identifiable up to the usual permutation and scaling ambiguities. However, in real bulk RNA-seq data, the observed cell-type or cell-stage proportions are often sparse, highly unbalanced, or clustered near only a subset of the simplex. For example, most observed samples may be dominated by the ring stage, trophozoite stage, or schizont stage. In this situation, the observed data alone may not provide enough geometric coverage, leaving the factorization ill-posed and unstable.

To address this limitation, GSNMF+ introduces a data augmentation strategy that supplements the real bulk samples with synthetically generated pseudo-bulk mixtures of known composition. These augmented samples are constructed to span the stage-composition simplex more broadly and therefore improve the practical identifiability of the NMF solution. For each augmented sample, a stage-composition vector is first generated from a Dirichlet distribution, $\mathbf{p} \sim \text{Dirichlet}(\boldsymbol{\alpha})$ where $\mathbf{p}$ denotes the known stage-proportion vector of the augmented sample, and $\boldsymbol{\alpha}$ controls the concentration of the sampled proportions.

To span the simplex broadly, we combine stage-dominant draws (small α concentrated on one stage) with near-uniform draws ($\boldsymbol{\alpha} = \mathbf{1}$); the exact concentration parameters used are reported with each experiment. Augmented samples are generated from one or more externally independent scRNA-seq reference datasets, producing N synthetic samples that are concatenated with the m real samples to form the com-

bined matrix $\tilde{\mathbf{G}} \in \mathbb{R}_{+}^{N\times(m+M)}$, used as input to all subsequent steps. The augmented samples serve two roles. First, by spanning the simplex more uniformly, they restore the scatter condition above and stabilize the factorization even when the real data alone are insufficient. Second, their known ground-truth proportions enable the component-annotation procedure described in Section (2.2.2) .

2.2.2 Estimated Component Annotation

A remaining ambiguity in NMF-based deconvolution is the permutation ambiguity of the estimated components. After factorization, the estimated components are not automatically ordered according to biologically meaningful stages. For example, the first estimated component may correspond to the ring stage in one run, but to the trophozoite or schizont stage in another run. Therefore, an annotation step is required to map each estimated NMF component to its corresponding biological cell stage.

GSNMF+ uses the augmented pseudo-bulk samples to resolve this ambiguity. Since the augmented samples are generated with known stage proportions, we can compare the estimated proportions from GSNMF+ with the known ground-truth proportions of the augmented samples. Specifically, after deconvolution, we extract the estimated component proportions for the augmented samples and compute their correlations with the known ring, trophozoite, and schizont proportions. This produces a correlation table between estimated components and true biological stages.

The final annotation is obtained by assigning each estimated component to the stage with which it has the strongest overall agreement. To ensure a one-to-one mapping, each estimated component is matched to only one biological stage, and each biological stage is assigned to only one estimated component. In implementation, this one-to-one assignment is solved as a linear sum assignment problem using the `solve_LSAP()` function from the R package `clue`. Before applying `solve_LSAP()`, the correlation matrix is shifted to be nonnegative by subtracting its minimum value. This transformation does not change the optimal matching, but makes the assignment step

computationally convenient. The assignment is then performed by maximizing the total correlation between the estimated components and the known stage proportions.

After the optimal matching is obtained, the estimated components are relabeled as ring, trophozoite, and schizont. This annotation step resolves the permutation ambiguity of NMF and makes the deconvolution output directly interpretable. It also allows the estimated proportions to be consistently compared with ground truth in simulation studies and across different reference datasets in real-data applications.

2.2.3 Multiplicative-Update-Based Optimization

The third improvement in GSNMF+ is the use of a multiplicative-update-based optimization procedure. The original GS-NMF framework was solved using the alternating direction method of multipliers (ADMM [59]), which is flexible but requires auxiliary variables and additional tuning parameters. In contrast, GSNMF+ adopts a multiplicative-update rule analogous to the Lee–Seung update scheme for NMF [60]. This approach is more straightforward to implement and naturally preserves the nonnegativity of both the expression matrix (C) and the proportion matrix (P) without requiring an additional projection step.

At each iteration, the update rules are derived from the gradient of the objective function. The gradient is decomposed into positive and negative components: terms that decrease the objective are collected in the numerator, while terms that increase the objective are collected in the denominator. The corresponding element-wise ratio updates ensure that all entries remain nonnegative throughout the optimization process. In practice, this produces stable convergence with relatively low computational and implementation complexity.

The full derivation of the multiplicative update rules, including the solvability constraint $\mathcal{F}_1$ and the manifold constraint $\mathcal{F}_2$, is provided in the Appendix A. The complete GSNMF+ procedure is summarized in Algorithm 1. The algorithm takes as input the combined expression matrix , It takes as input the combined matrix

$\tilde{\mathbf{G}}$, initial factors $\mathbf{C}_0, \mathbf{P}_0$, and the regularization parameters λ_s, λ_m. Overall, replacing ADMM with multiplicative updates simplifies the optimization procedure while retaining the geometric constraints that improve identifiability and interpretability.

Algorithm 1 GSNMF+

Require: Augmented data matrix $\tilde{\mathbf{G}}$; initial values $\mathbf{C}_0$, $\mathbf{P}_0$; structure labels $\mathbf{C}_{0,\text{labels}}$; tolerance ε; maximum iterations T; parameters λ_s, λ_m

1: Initialize $\mathbf{C} \leftarrow \mathbf{C}_0$, $\mathbf{P} \leftarrow \mathbf{P}_0$
2: Compute marker-structure matrices W_1, W_2, C_g
3: Compute manifold weights W_3, W_4
4: **repeat**
5: Update $\mathbf{C}$:

$$\mathbf{C}^{k+1} \leftarrow \mathbf{C}^k \odot \frac{\mathbf{G}(\mathbf{P}^k)^\top + \lambda_s W_1^k \mathbf{C}_{\text{ref}} + \lambda_m W_4^k \mathbf{C}^k}{\mathbf{C}^k \mathbf{P}^k (\mathbf{P}^k)^\top + \lambda_s W_1^k W_2^k \mathbf{C}^k + \lambda_m \operatorname{diag}(W_3^k \mathbf{1}) \operatorname{diag}(|\mathbf{C}^k|^{-2})\, \mathbf{C}^k}$$

6: Update $\mathbf{P}$:

$$\mathbf{P}^{k+1} \leftarrow \mathbf{P}^k \odot \frac{(\mathbf{C}^{k+1})^\top \tilde{\mathbf{G}}}{(\mathbf{C}^{k+1})^\top \mathbf{C}^{k+1} \mathbf{P}^k}$$

7: Compute loss $\mathcal{L} = \frac{1}{2}\|\tilde{\mathbf{G}} - \mathbf{C}^{k+1}\mathbf{P}^{k+1}\|_F^2 + \lambda_s \mathcal{F}_1(\mathbf{C}^{k+1}) + \lambda_m \mathcal{F}_2(\mathbf{C}^{k+1})$
8: **until** relative change in both $\mathbf{C}$ and $\mathbf{P}$ falls below ε, or a maximum of $T = 2,000$ iterations reached
9: **return** $\mathbf{C}$, $\mathbf{P}$

The model involves two tunable regularization parameters, λ_s (solvability) and λ_m (manifold); the manifold bandwidth σ is held fixed (Section above) and is not part of the search. The two λ parameters are selected by grid search over the candidate set $\Lambda = \{0.01, 0.06, 0.11, \ldots, 0.96, 1.0, 1.5, 2.0, \ldots, 10.0\}$. For each pair $(\lambda_s, \lambda_m) \in \Lambda \times \Lambda$, the model is fitted with the same initialization $(\mathbf{C}_0, \mathbf{P}_0)$ and marker genes, and iterated until convergence or the iteration cap. Convergence is assessed by the relative change in the normalized Frobenius norms of $\mathbf{C}$ and $\mathbf{P}$ between successive iterations. The pair minimizing the final objective value is selected.

After introducing the three enhancements to the baseline NMF framework, we summarize the complete workflow of the proposed GSNMF+ method in Figure Fig2.3

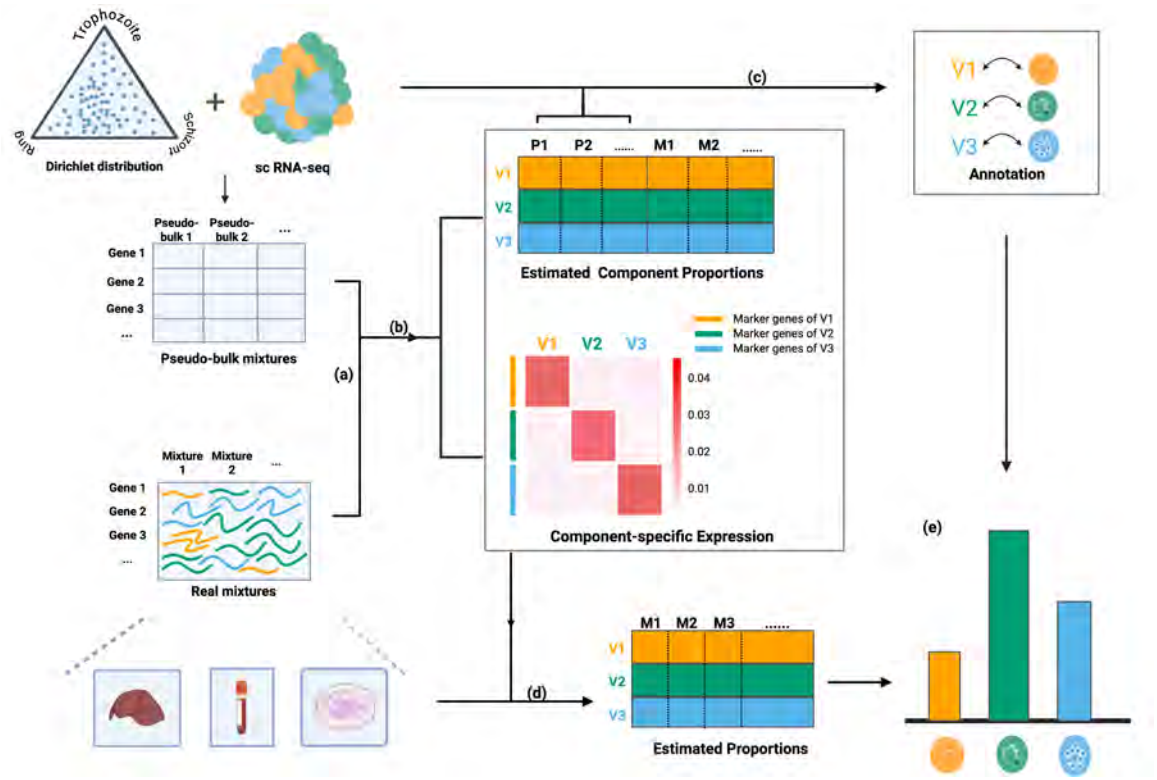


Figure 2.3: GSNMF+ workflow

Modules: (a) Generate pseudo-bulk mixtures and combine with real bulk tissue data. (b)Apply GSNMF to the combined dataset for deconvolution. (c) Component annotation using pseudo-bulk mixtures' cell proportion information. (d) Apply GSNMF to the real bulk tissue dataset. (e) Annotated cell composition estimation for real bulk tissue sample.

2.3 Data Description

2.3.1 Reference Single-Cell Datasets

Three publicly available single-cell *Plasmodium* transcriptomic datasets were used as reference data in this study (Table 2.1). The first dataset, generated by Howick et al. [13], was derived from in vitro cultures of the *P. falciparum* Pf3D7 strain using Illumina sequencing. After quality control, ortholog mapping, and expression filtering, where genes expressed in at least 100 cells were retained, this dataset contained

3,595 genes across 6,493 cells. The cells covered three intraerythrocytic developmental stages: ring (n=1,251), trophozoite (n=4,477), and schizont (n=765). The second dataset was also generated by Howick et al. [13], but from in vitro cultures of the *P. knowlesi* PKNH strain. After the same preprocessing procedure, this dataset contained 2,746 genes across 4,230 cells, including ring (n=790), trophozoite (n=1,793), and schizont (n=1,647) stages. The third dataset, generated by Dogga et al. [58], was derived from *P. falciparum* Pf3D7 parasites. After applying the same preprocessing steps, the dataset contained 4,923 genes across 29,216 cells, including ring (n=7,499), trophozoite (n=13,558), and schizont (n=8,159) stages.

Table 2.1: Information about the three Malaria single cell RNA sequencing datasets used in this study.

Dataset	**Source**	**Strain**	**No. of genes**	**No. of cells (at stages)**
Howick-pf [13]	Laboratory	Pf3D7	3,595	6,493 (R:1,251, T:4,477, S:765)
Howick-pk [13]	Laboratory	PKNH	2,746	4,230 (R:790, T:1,793, S:1,647)
Dogga-pf [58]	Laboratory	Pf3D7	4,923	29,216 (R:7,499, T:13,558, S:8,159)

R = Ring, T = Trophozoite, S = Schizont.

2.3.2 Real Bulk Mixture Datasets

Two real bulk RNA-seq datasets from malaria transcriptomic studies were used for real-data evaluation. Both datasets contain mixed intraerythrocytic parasite stages and therefore provide practical examples for assessing the performance of GSNMF+ on malaria bulk mixture deconvolution. Since ground-truth stage proportions are unavailable for these real samples, the analysis focuses on the consistency and biological interpretability of the estimated ring, trophozoite, and schizont proportions. The processed datasets are summarized in (Table 2.2).

Table 2.2: Summary of the real bulk RNA-seq datasets after quality control (QC), ortholog mapping, and preprocessing. Genes expressed in at least 10 samples (mixture_Kim) or 3 samples (mixture_Kepple) were retained for analysis.

Mixture	**Source**	**Strain**	**No. of genes**	**No. of samples**
Kim [61]	Laboratory	PvP01	3682	26
Kepple [62]	Laboratory	PvP01	3785	10

2.3.3 Type 2 Diabetes Datasets

To evaluate whether GSNMF+ can be generalized beyond malaria bulk RNA-seq deconvolution, we also applied the framework to type 2 diabetes (T2D) pancreatic transcriptomic datasets. As summarized in Table 2.3, two single-cell RNA-seq datasets were used as reference data, including GSE81608-Xin and E-MTAB-5061, both derived from human pancreas samples. The real bulk mixture dataset, GSE50244, was also generated from human pancreas tissue and contained 89 bulk samples. These datasets provide a separate disease context for assessing the applicability of GSNMF+ to heterogeneous human tissue data.

Table 2.3

Data	**Type**	**Tissue Type**	**No. of genes**	**No. of cells/samples**
GSE81608-xin [63]	SC	human pancreas	27,461	1,492
E-MTAB-5061 [64]	SC	human pancreas	20,444	748
GSE50244 [65]	Mixture	human pancreas	32,581	89

2.4 Simulation Results

To evaluate the performance of GSNMF+ under controlled settings, we constructed pseudo-bulk mixture datasets from single-cell *Plasmodium* reference data with known ground-truth stage proportions. The goal of the simulation study was to assess whether GSNMF+ can accurately recover the proportions of ring, trophozoite, and

schizont stages from mixed bulk-like expression profiles. In addition, the simulations were designed to examine the robustness of the proposed augmentation strategy under different stage-composition patterns, including balanced mixtures and stage-dominant mixtures.

The performance of GSNMF+ was compared with existing deconvolution methods, including CIBERSORTx[23] and BayesPrism [25]. Since the true stage proportions are known in the simulation setting, performance was evaluated using mean squared error (MSE) and Pearson correlation between the true and estimated proportions for each stage.

2.4.1 Pseudo-Bulk Mixture Construction

To evaluate deconvolution performance under controlled settings, pseudo-bulk mixtures were constructed from single-cell RNA-seq reference datasets. For each simulated pseudo-bulk sample, a cell-stage proportion vector $\boldsymbol{p}$ was first generated according to the simulation scenario described in the subsection 2.4.2. The vector ($\boldsymbol{p}$) specifies the relative proportions of ring, trophozoite, and schizont cells in the simulated mixture.

Given $\boldsymbol{p}$, 1,000 individual cells were randomly sampled with replacement from the corresponding single-cell RNA-seq dataset, stratified according to the sampled stage proportions. The raw read counts of the selected cells were then summed genewise to generate a single pseudo-bulk expression profile. This construction mimics the aggregation process observed in bulk RNA-seq experiments, where the measured expression profile represents a mixture of signals from multiple cell stages.

Because the pseudo-bulk mixtures are generated from known sampling proportions, the true cell-stage proportions are available for every simulated sample. This allows direct evaluation of deconvolution accuracy by comparing the estimated proportions with the ground-truth proportions. For each simulation scenario, 100 pseudo-bulk mixtures were generated as simulated bulk samples. In addition, 290 pseudo-bulk

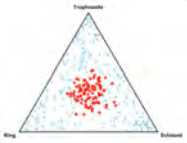
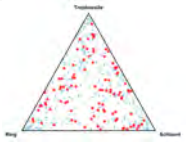
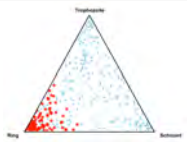
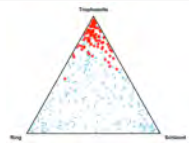
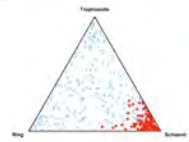
mixtures were generated as augmentation samples and combined with the simulated bulk data for GSNMF+.

2.4.2 Cell-Stage Proportion Generation

Cell-stage proportions for each pseudo-bulk sample were generated from a multivariate Dirichlet distribution. Specifically, for each simulated mixture, the stage-proportion vector was drawn as $\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\alpha})$, where $\boldsymbol{\pi} = (\pi_R, \pi_T, \pi_S)$ represents the proportions of ring, trophozoite, and schizont stages, respectively. The concentration parameter vector $\boldsymbol{\alpha}$ controls both the expected stage composition and the variability of the simulated mixtures. Larger relative values of a given component in $\boldsymbol{\alpha}$ generate mixtures enriched for the corresponding stage, while equal values across all components produce balanced mixtures. The total concentration $\sum_{\ell=1}^{k} \alpha_\ell$ controls the overall variability of the proportions, with larger values producing less variable mixtures.

Five simulation scenarios were defined to represent a range of biologically relevant cell-stage composition patterns. The first two scenarios were designed to generate balanced mixtures with different levels of compositional variation. In Scenario 1, proportions were drawn from $\text{Dirichlet}(10, 10, 10)$, producing approximately equal mean stage proportions with relatively low variability. In Scenario 2, proportions were drawn from $\text{Dirichlet}(1, 1, 1)$, producing equal mean stage proportions but higher variability across samples. The remaining three scenarios were designed to represent stage-dominant mixtures. Scenario 3 used $\text{Dirichlet}(10, 1, 1)$ to generate ring-dominant mixtures, Scenario 4 used $\text{Dirichlet}(1, 10, 1)$ to generate trophozoite-dominant mixtures, and Scenario 5 used $\text{Dirichlet}(1, 1, 10)$ to generate schizont-dominant mixtures. For each scenario, 100 pseudo-bulk samples were generated. The simulated cell-stage compositions of these pseudo-bulk samples are shown as red points in the simplex plots in Table 2.4.

Table 2.4: Cellular compositions shown as simplexes. Red dots represent cellular compositions of the 100 pseudo-bulk samples, and blue points represent the cellular composition of the 290 augmentation samples.

S1	S2	S3	S4	S5
				

In addition to the simulated pseudo-bulk samples, a fixed augmentation dataset containing 290 samples was generated and used across all scenarios. This augmentation dataset was designed to improve simplex coverage and support stable deconvolution. It consisted of 90 stage-dominant mixtures, with 30 samples each generated from Dirichlet$(10, 1, 1)$, Dirichlet$(1, 10, 1)$, and Dirichlet$(1, 1, 10)$, together with 200 high-variation balanced mixtures generated from Dirichlet$(1, 1, 1)$. The cell-stage compositions of the augmentation samples are shown as blue points in the simplex plots in Table 2.4.

2.4.3 Simulation Results

Using the pseudo-bulk mixtures and cell-stage proportion settings described above, we evaluated whether GSNMF+ could accurately recover the known ring, trophozoite, and schizont proportions under different compositional regimes. The simulation design provides a controlled benchmark because the true stage proportions are known for every pseudo-bulk sample.

To mimic the practical setting in which reference datasets are often generated from independent studies, the augmentation samples were constructed from an external single-cell RNA-seq reference dataset rather than from the same dataset used to generate the simulated pseudo-bulk mixtures. In this study, we considered two simulation cases. In Case 1, pseudo-bulk mixtures were generated from the Howick *P. knowlesi* reference dataset, while the augmentation samples were generated from the Dogga *P.*

falciparum reference dataset. This case represents a cross-species setting, where the target pseudo-bulk data and the augmentation reference come from different *Plasmodium* species. In Case 2, pseudo-bulk mixtures were generated from the Howick *P. falciparum* reference dataset, while the augmentation samples were again generated from the Dogga *P. falciparum* reference dataset. This case represents a within-species but cross-study setting, where both datasets are from *P. falciparum* but differ in experimental source and reference composition. Together, these two cases allow us to evaluate whether GSNMF+ remains robust when the augmented reference data and the target pseudo-bulk data are not perfectly matched.

We compared GSNMF+ with two widely used reference-based deconvolution methods, CIBERSORTx [23] and BayesPrism [25]. For each method, deconvolution accuracy was evaluated using mean squared error (MSE) and Pearson correlation between the estimated and true stage proportions. MSE was used to quantify absolute estimation error, while Pearson correlation was used to assess whether the estimated proportions preserved the relative variation across samples. Both metrics were computed separately for the ring, trophozoite, and schizont stages across the five simulation scenarios.

2.4.3.1 Case 1: Howick-*pk* Pseudo-bulk Data with Dogga-*pf* Augmentation

In the first simulation case, the Howick-*pk* single-cell dataset was used to generate the pseudo-bulk mixtures, while the Dogga-*pf* dataset was used to generate the augmentation samples. This setting represents a cross-species reference scenario, where the target pseudo-bulk mixtures and the augmentation reference are derived from different *Plasmodium* species. Therefore, this case provides a challenging setting for evaluating whether GSNMF+ can remain accurate when the reference data used for augmentation are not perfectly matched to the target mixtures.

We compared the deconvolution performance of GSNMF+ with two reference-based methods, BayesPrism[25] and CIBERSORTx[23]. For each method, the estimated

stage proportions were compared with the known ground-truth proportions of the simulated pseudo-bulk samples. Performance was evaluated using mean squared error (MSE) and true-versus-estimated proportion scatter plots across the five simulation scenarios.

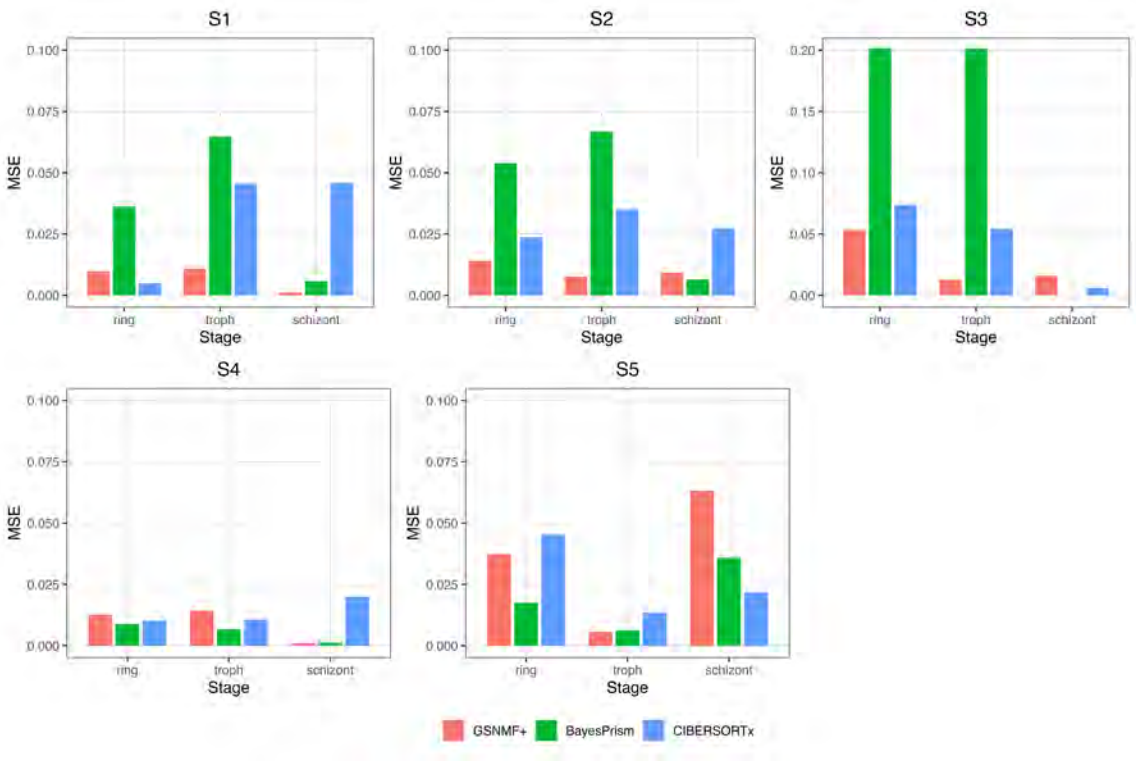


Figure 2.4: MSE comparison for Case 1, where pseudo-bulk mixtures were generated from the Howick-*pk* dataset and augmentation samples were generated from the Dogga-*pf* dataset. MSE was computed between the true and estimated proportions for each parasite stage across five simulation scenarios.

As shown in Figure 2.4, GSNMF+ generally achieved low MSE across most scenarios and stages. In the balanced scenarios S1 and S2, GSNMF+ produced smaller errors than the comparison methods for most stages, indicating stable performance when the true stage proportions were approximately balanced. In the stage-dominant scenarios, GSNMF+ also remained robust overall. Although its largest error occurred in S5, the schizont-dominant setting, for the schizont stage, the MSE remained below 0.07, suggesting that the method retained good accuracy even in this challenging set-

ting. In contrast, the comparison methods showed larger and less consistent errors in several scenarios. BayesPrism had substantially larger errors in S3 for the ring and trophozoite stages, while CIBERSORTx showed relatively large errors in the balanced scenarios, especially for the trophozoite and schizont stages in S1 and S2.

The true-versus-estimated scatter plots in Figure 2.5 further support the MSE results and reveal the bias patterns of each method. GSNMF+ produced the most compact and diagonal-aligned estimates across most scenarios, indicating strong agreement between estimated and true proportions. In S1 and S3, the estimates still showed some stage-dependent compression or offset from the diagonal, suggesting that systematic bias can remain under cross-species mismatch. However, the overall point patterns were more structured and less dispersed than those of the comparison methods.

BayesPrism tracked the diagonal reasonably well in several scenarios, but showed clear deviations in others. In particular, the S3 scatter plots showed substantial over- or under-estimation for some stages, consistent with the high MSE observed in Figure 2.4. CIBERSORTx showed the largest dispersion overall, and in several scenarios some estimated proportions were close to zero even when the true proportions were nonzero. This pattern suggests that CIBERSORTx may fail to recover under-represented stages in some mixtures.

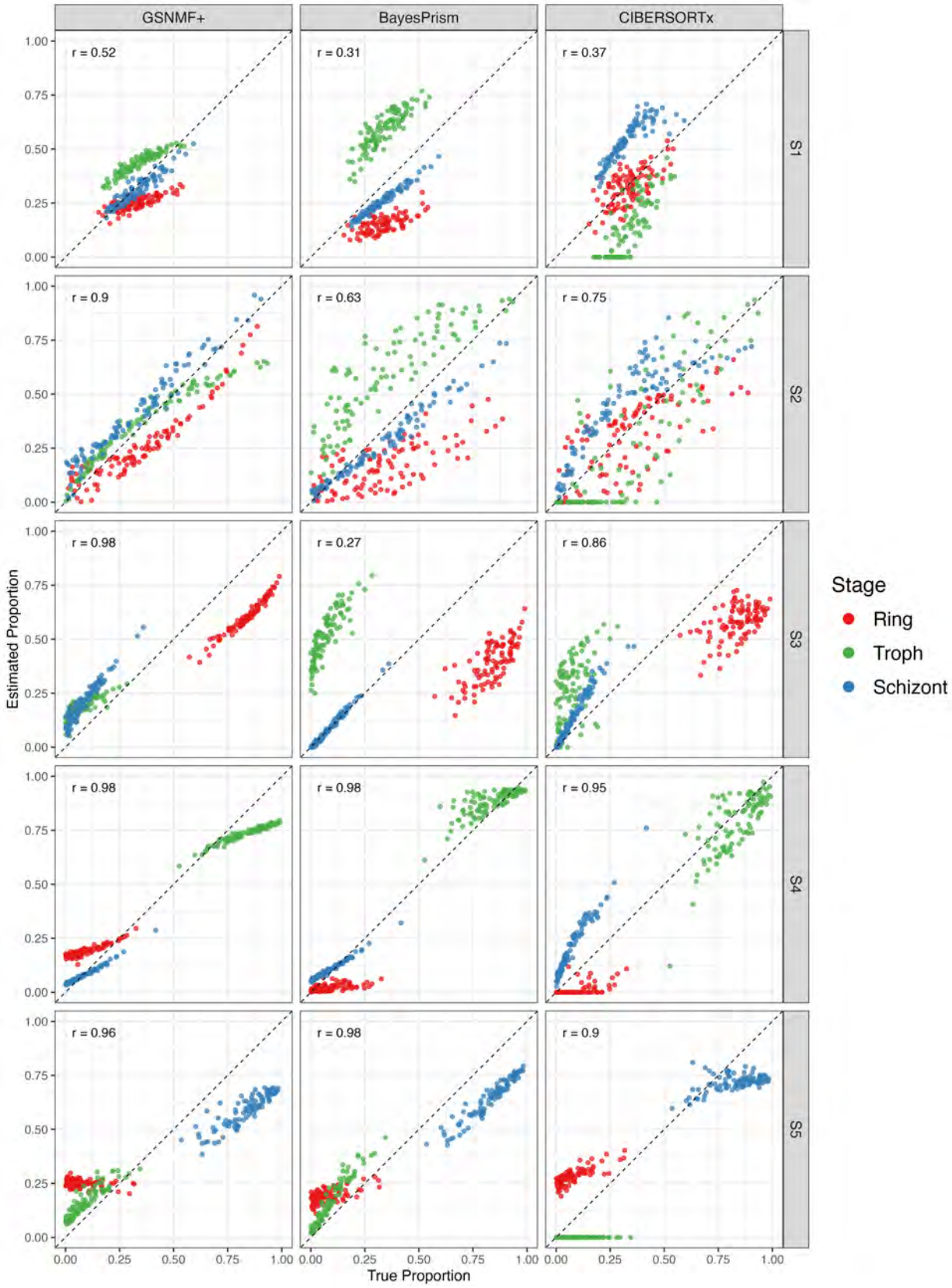


Figure 2.5: True versus estimated stage proportions for Case 1. Each point represents one simulated pseudo-bulk sample, and the dashed diagonal line indicates perfect agreement between the true and estimated proportions. Columns correspond to deconvolution methods, and rows correspond to simulation scenarios.

Overall, Case 1 demonstrates that GSNMF+ can maintain relatively accurate and stable deconvolution performance even under a cross-species simulation setting. The improvement is especially evident in scenarios where the pseudo-bulk mixtures are balanced or where non-dominant stages need to be recovered. These results suggest that the augmentation strategy and component annotation procedure help improve robustness when the augmentation reference and target pseudo-bulk data are not perfectly matched.

2.4.3.2 Case 2: Howick-*pf* Pseudo-bulk Data with Dogga-*pf* Augmentation

In the second simulation case, the Howick-*pf* single-cell dataset was used to generate the pseudo-bulk mixtures, while the Dogga-*pf* dataset was used to generate the augmentation samples. Unlike Case 1, both datasets are from *P. falciparum*; however, they were generated from different studies and have different single-cell expression characteristics. Therefore, this case represents a within-species but cross-study setting, allowing us to evaluate whether GSNMF+ remains robust when the pseudo-bulk data and augmentation reference come from different *P. falciparum* datasets.

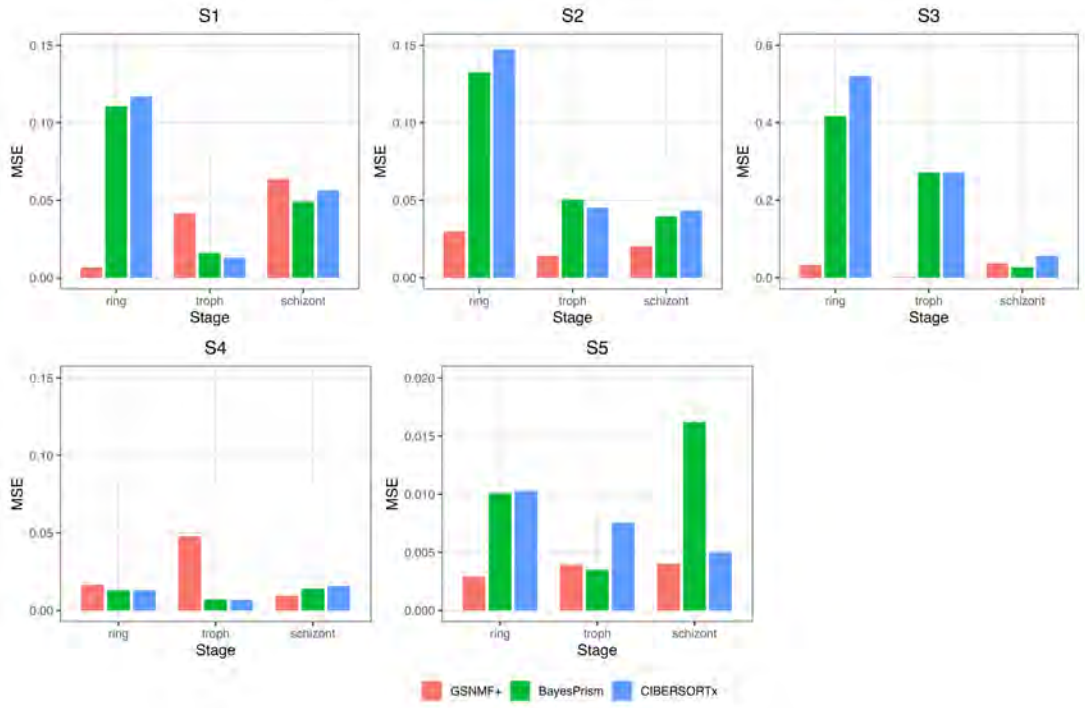


Figure 2.6: MSE comparison for Case 2, where pseudo-bulk mixtures were generated from the Howick-*pf* dataset and augmentation samples were generated from the Dogga-*pf* dataset. MSE was computed between the true and estimated proportions for each parasite stage across five simulation scenarios.

As shown in Figure 2.6, GSNMF+ achieved low MSE in most scenarios, especially in S2, S3, and S5. In S2, where the true stage proportions were balanced but highly variable, GSNMF+ produced the smallest errors across all three stages. In S3, the ring-dominant scenario, GSNMF+ substantially reduced the estimation error compared with BayesPrism and CIBERSORTx, particularly for the ring and trophozoite stages, where the comparison methods showed large errors. In S5, the schizont-dominant scenario, GSNMF+ also maintained very small errors across all stages. In S1 and S4, GSNMF+ showed slightly higher errors for some stages, particularly the schizont stage in S1 and the trophozoite stage in S4. However, these increases were relatively small, and the overall MSE remained within a moderate range, indicating that GSNMF+ still provided stable deconvolution performance across the full set of simulation scenarios.

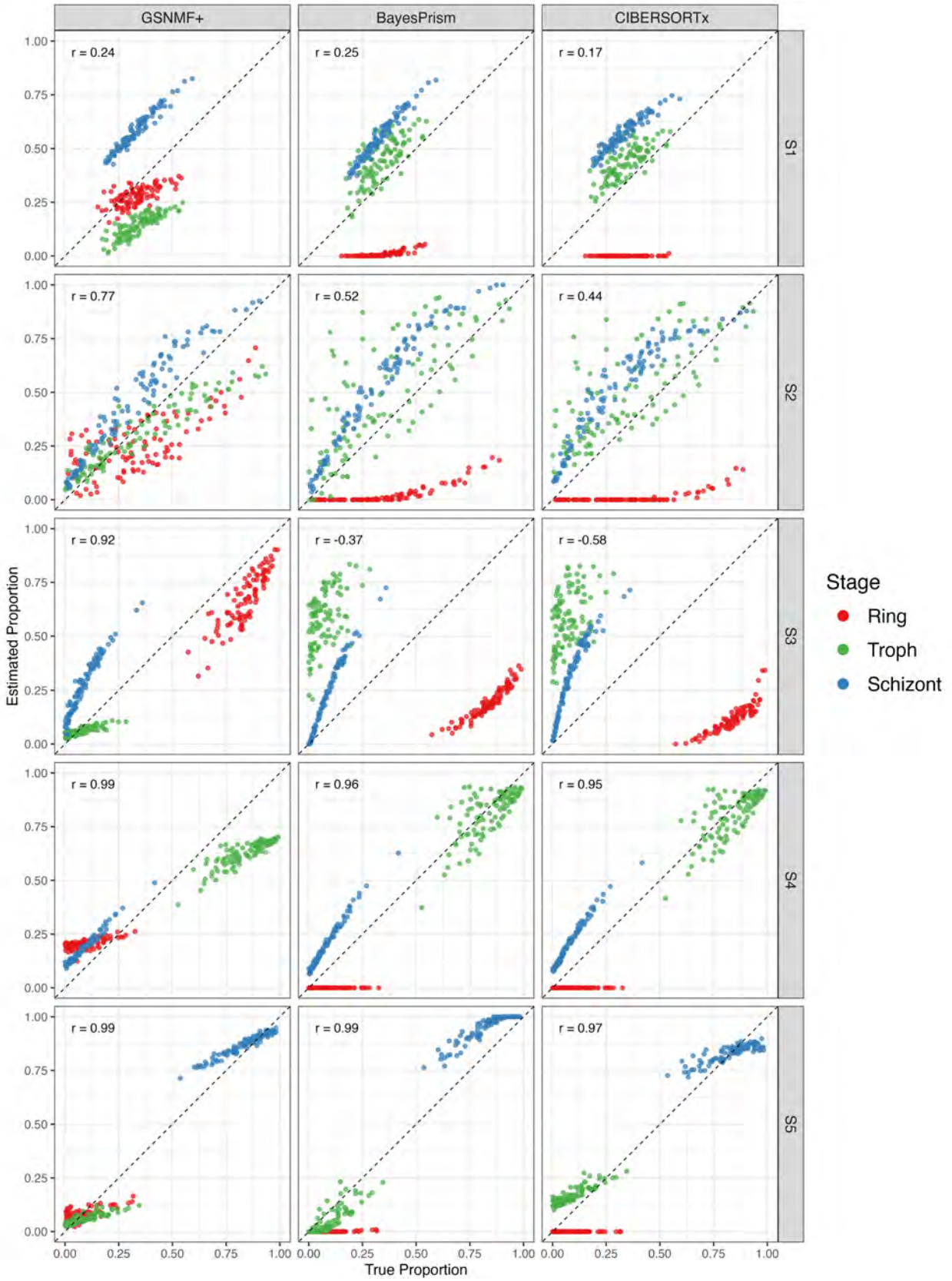


Figure 2.7: True versus estimated stage proportions for Case 2. Each point represents one simulated pseudo-bulk sample, and the dashed diagonal line indicates perfect agreement between the true and estimated proportions. Columns correspond to deconvolution methods, and rows correspond to simulation scenarios.

The true-versus-estimated scatter plots in Figure 2.7 provide a more detailed view of the estimation patterns. For GSNMF+, the estimates were strongly aligned with the diagonal in S2, S3, S4, and S5, indicating that the method preserved the relative variation of the true stage proportions in these scenarios. The strongest agreement was observed in the stage-dominant scenarios, especially S4 and S5, where the overall Pearson correlations were close to one.

The comparison methods showed more pronounced artifacts in several scenarios. In particular, BayesPrism and CIBERSORTx frequently estimated the ring-stage proportion as zero or near zero in multiple scenarios, as shown by the red points concentrated along the bottom axis. This behavior explains their larger ring-stage MSE in S1–S3. In contrast, GSNMF+ avoided this near-zero ring-stage artifact and produced more stable estimates for the under-represented stages.

Overall, Case 2 shows that GSNMF+ remains effective when both pseudo-bulk and augmentation data are derived from *P. falciparum* but from different studies. The results suggest that the augmentation strategy improves robustness to cross-study reference differences and helps stabilize stage-proportion estimation, particularly in high-variation and stage-dominant simulation settings.

2.5 Stability

In addition to the main simulation comparison, we further evaluated the robustness of GSNMF+ using repeated experiments. For each simulation scenario, pseudo-bulk mixtures were regenerated 100 times under the same scenario-specific cell-stage proportion distribution, and deconvolution was performed independently for each replicate. Robustness was summarized using the median MSE across the 100 repeated experiments for each method, stage, and scenario. Because CIBERSORTx is run through a web interface, performing 100 repeated experiments for each scenario was impractical. Therefore, this robustness analysis compares GSNMF+ with BayesPrism.

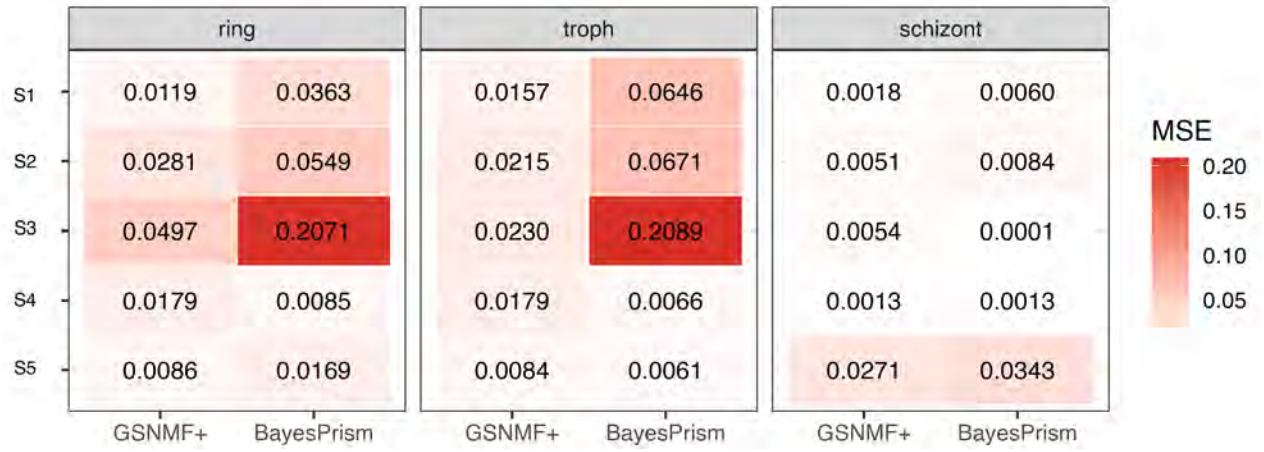


Figure 2.8: Robustness analysis for Case 1, where pseudo-bulk mixtures were generated from the Howick-*pk* dataset and augmentation samples were generated from the Dogga-*pf* dataset. Each cell shows the median MSE across 100 repeated experiments.

Figure 2.8 shows the robustness results for Case 1. Overall, GSNMF+ achieved lower median MSE than BayesPrism in most scenarios and stages. The improvement was most pronounced in the ring-dominant scenario S3. In this scenario, GSNMF+ obtained median MSE values of 0.0497 for the ring stage and 0.0230 for the trophozoite stage, compared with 0.2071 and 0.2089 for BayesPrism, respectively. This indicates that GSNMF+ was substantially more stable when the mixture was dominated by the ring stage. GSNMF+ also showed consistently lower errors across S1 and S2 for all three stages. BayesPrism produced slightly lower errors only in a few low-error settings, including the schizont stage in S3 and the ring and trophozoite stages in S4. For most other settings, especially those involving ring-stage estimation, GSNMF+ showed stronger robustness.

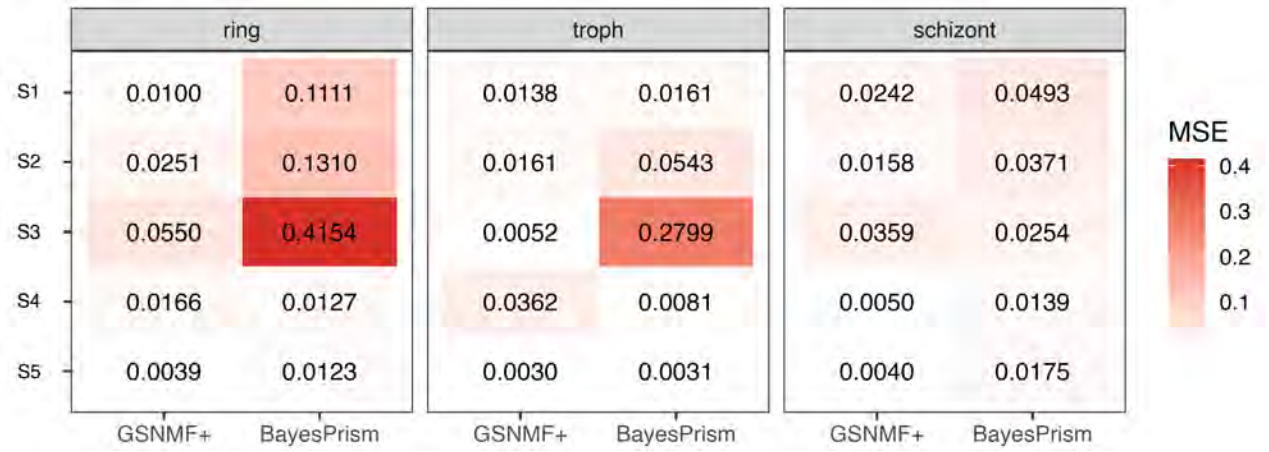


Figure 2.9: Robustness analysis for Case 2, where pseudo-bulk mixtures were generated from the Howick-*pf* dataset and augmentation samples were generated from the Dogga-*pf* dataset. Each cell shows the median MSE across 100 repeated experiments.

A similar pattern was observed in Case 2 (Figure 2.9), where GSNMF+ showed clear advantages over BayesPrism, especially for ring-stage estimation. In S1–S3, the median MSE values for BayesPrism at the ring stage were 0.1111, 0.1310, and 0.4154, respectively, whereas the corresponding values for GSNMF+ were much lower, at 0.0100, 0.0251, and 0.0550. The largest improvement again occurred in S3, the ring-dominant scenario. In this setting, GSNMF+ substantially reduced the ring-stage error and also achieved a much smaller trophozoite-stage MSE than BayesPrism, with values of 0.0052 versus 0.2799. These results indicate that GSNMF+ provides more robust stage-proportion recovery under strong composition imbalance. Although BayesPrism had lower MSE in a few isolated settings, such as the trophozoite stage in S4, GSNMF+ achieved lower errors in the majority of scenario-stage combinations.

Taken together, the repeated simulation experiments demonstrate that GSNMF+ provides more stable deconvolution performance across both simulation cases. Its largest gains occur in the ring-dominant scenario and for ring-stage estimation more generally, consistent with the patterns observed in the main MSE and scatter-plot analyses. These results support the robustness of GSNMF+ when pseudo-bulk sam-

ples are compositionally imbalanced or when one stage is difficult to recover.

2.6 Real Data Application

After evaluating GSNMF+ on simulated pseudo-bulk mixtures with known ground-truth proportions, we next applied the method to real bulk RNA-seq datasets. In real data applications, the true underlying cell-type or cell-stage proportions are generally unknown, so direct accuracy evaluation based on MSE or correlation with ground truth is not possible. Therefore, we evaluated method performance by examining the consistency of estimated proportions across independent single-cell reference datasets. In practice, reference datasets are often generated from separate studies and may differ in species, experimental platform, sequencing depth, or cell-state composition. A reliable deconvolution method should therefore produce broadly concordant estimates when different biologically relevant references are used, whereas strong disagreement across references indicates sensitivity to reference choice.

We applied this evaluation to three real bulk RNA-seq datasets spanning two biological systems. The first two datasets are malaria bulk RNA-seq datasets from *P. vivax*: the Kim dataset [61] and the Kepple dataset [62]. For these datasets, we decomposed each bulk sample into three intraerythrocytic developmental stages: ring, trophozoite, and schizont. To assess reference robustness, we used three referencence, including two *P. falciparum* single-cell references, Dogga-*pf* [58] and Howick-*pf* [13], together with another reference, Howick-*pk* [13]. This design allowed us to evaluate reference variability and cross-species reference mismatch.

The third real dataset is a human pancreatic islet bulk RNA-seq dataset from donors with and without type 2 diabetes [65]. This dataset was used to evaluate whether GSNMF+ can generalize beyond malaria transcriptomic deconvolution. For this application, bulk samples were decomposed into four endocrine cell types using two independent single-cell pancreas references [63, 64]. In addition to reference consistency, the disease-status labels provide an external biological context for assessing

whether the estimated cell-type proportions are biologically plausible.

For all real-data analyses, batch effects between real bulk samples and augmented pseudo-bulk samples were addressed before deconvolution. Specifically, ComBat, implemented in the `sva` package, was used for batch-effect correction [66, 67]. This step was applied to reduce systematic differences introduced by different data sources, sequencing platforms, and sample construction procedures. The real bulk samples and augmented pseudo-bulk samples were treated as separate batches, and the corrected expression matrix was then used as input for GSNMF+ and downstream comparison. This preprocessing step was included to improve comparability between the real and augmented data while preserving the biological variation needed for deconvolution.

Across all real-data applications, GSNMF+ was benchmarked against established reference-based deconvolution methods under matched preprocessing and reference settings. Because the true proportions are unavailable, the real-data analysis focuses on reference consistency, biological interpretability, and robustness of the estimated cellular compositions.

2.6.1 Mixture-Kim Dataset

We first evaluated the real-data performance of GSNMF+ on the mixture-Kim malaria bulk RNA-seq dataset. Since the true stage proportions of these real bulk samples are unknown, the goal of this analysis was not to compute prediction error, but to assess whether the estimated proportions remained consistent when different single-cell reference datasets were used. Specifically, we applied each method using three reference datasets: Dogga-*pf*, Howick-*pf*, and Howick-*pk*. The Dogga-*pf* and Howick-*pf* references provide a within-species comparison, while Howick-*pk* provides a cross-species comparison.

Figure 2.10 shows the pairwise comparison of estimated stage proportions obtained from the three references. For GSNMF+, the estimated proportions were broadly consistent across references. The overall pairwise correlations were 0.794 between

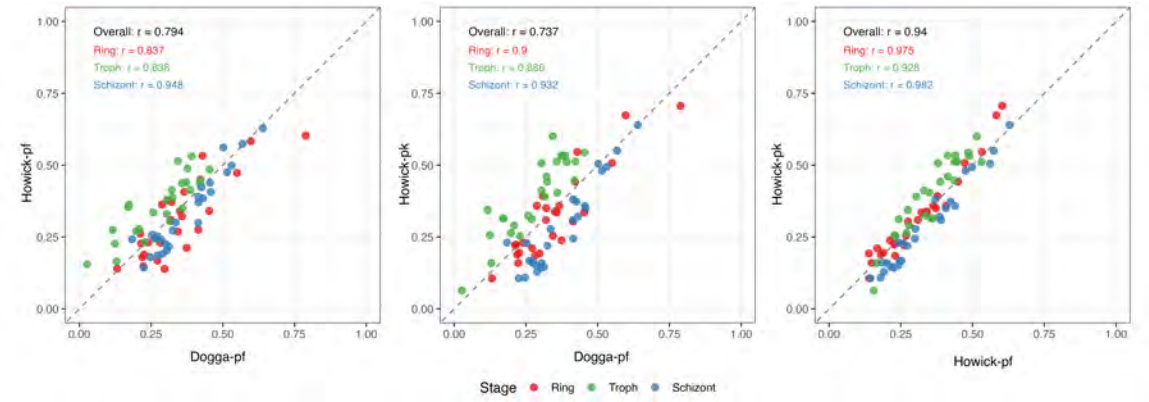
Dogga-*pf* and Howick-*pf*, 0.737 between Dogga-*pf* and Howick-*pk*, and 0.940 between Howick-*pf* and Howick-*pk*. The stage-specific correlations were also high, especially for the schizont stage, suggesting that GSNMF+ produced stable estimates even when the reference dataset changed. This indicates that the proposed augmentation and component-annotation strategy can reduce sensitivity to reference choice in real malaria bulk data.

In comparison, BayesPrism and CIBERSORTx showed stronger reference-dependent behavior. Although BayesPrism achieved high agreement in some reference pairs, its estimates showed larger discrepancies for the ring and trophozoite stages, especially when Dogga-*pf* was compared with the other references. CIBERSORTx showed the most reference-dependent estimates overall, particularly for the schizont stage, where the agreement varied substantially across reference pairs. These patterns suggest that the estimated proportions from the comparison methods were more sensitive to the selected single-cell reference dataset.

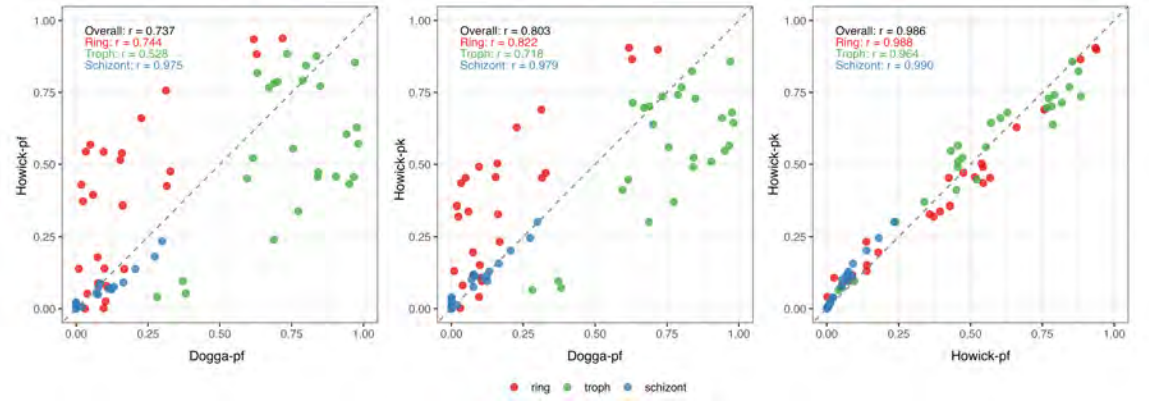
Figure 2.11 further illustrates the estimated proportions across individual bulk samples. For GSNMF+, the three reference-based trajectories followed similar sample-level patterns within each stage. The lines from Dogga-*pf*, Howick-*pf*, and Howick-*pk* largely overlapped, indicating that GSNMF+ produced stable stage-proportion estimates across samples. In contrast, BayesPrism and CIBERSORTx showed greater separation among the reference-specific trajectories, especially for the ring and trophozoite stages. This separation indicates that their estimated proportions depended more strongly on the chosen reference.

Overall, the mixture-Kim analysis shows that GSNMF+ provides more reference-consistent estimates than the comparison methods on real malaria bulk RNA-seq data. Even when a cross-species reference was included, GSNMF+ maintained broadly concordant estimates of ring, trophozoite, and schizont proportions. These results support the robustness of GSNMF+ in real-data settings where ground-truth proportions

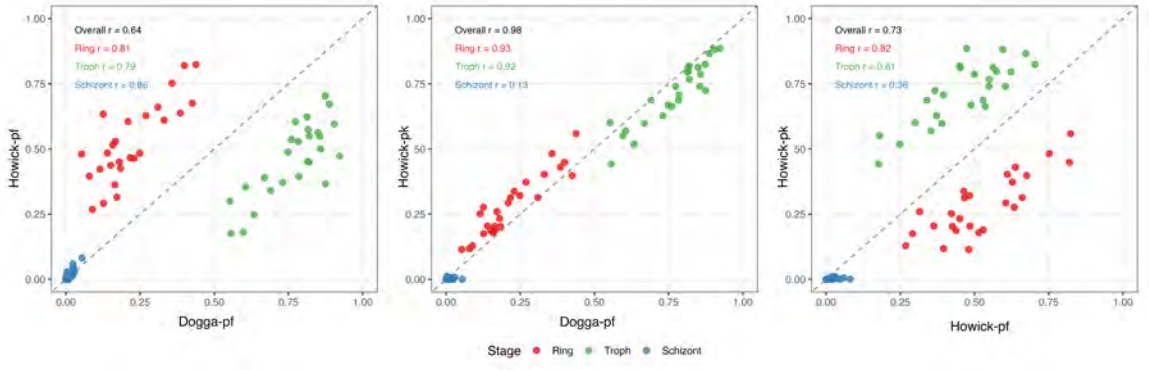
are unavailable and reference mismatch is unavoidable.



(a)

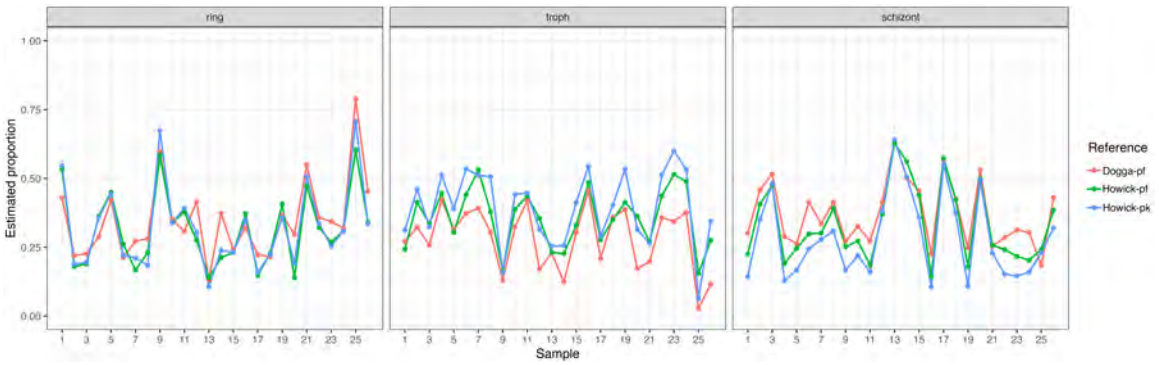


(b)

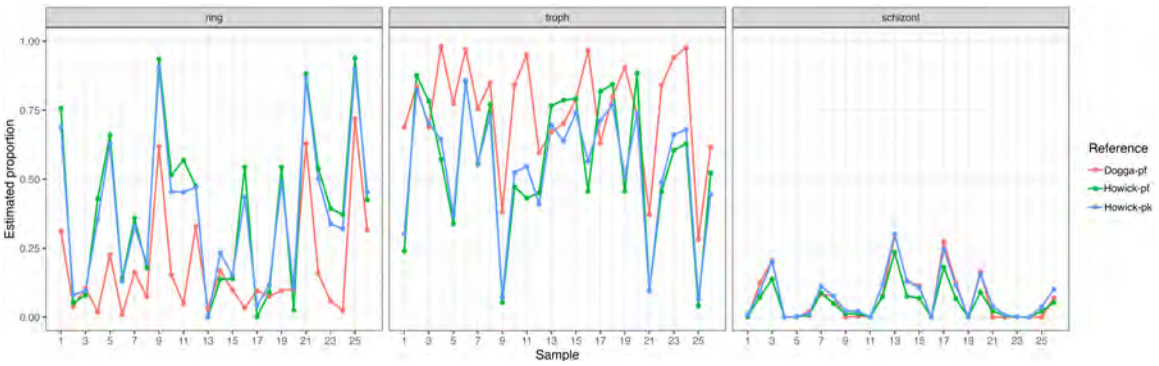


(c)

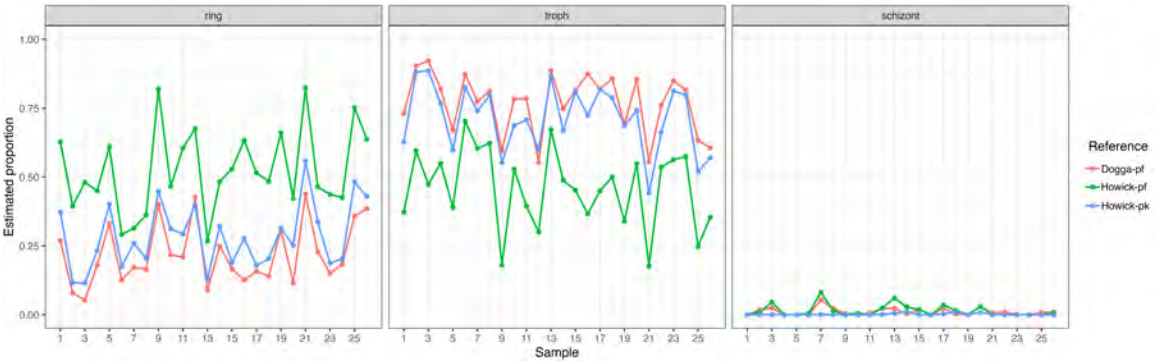
Figure 2.10: Pairwise comparison of estimated stage proportions obtained using the Dogga-pf, Howick-pf, and Howick-pk reference datasets for the mixture-Kim dataset. Colors denote Ring, Trophozoite, and Schizont stages. The dashed line indicates perfect agreement. (a) GSNMF+, (b) BayesPrism, (c) CIBERSORTx.



(a)



(b)



(c)

Figure 2.11: Estimated stage proportions across samples obtained using the Dogga-pf, Howick-pf, and Howick-pk reference datasets for the mixture-Kim dataset. Each panel corresponds to a parasite stage (ring, trophozoite, schizont), and within each panel the estimated proportion is plotted for every sample, with line color denoting the reference dataset used (Dogga, red; Howick-pf, green; Howick-pk, blue). Overlapping lines indicate agreements, whereas separation indicates that the estimated proportion depends on the choice of reference. (a) GSNMF+, (b) BayesPrism, and (c) CIBERSORTx

2.6.2 Mixture-Kepple Dataset

We also applied the same reference-consistency analysis to the mixture-Kepple malaria bulk RNA-seq dataset. As in the mixture-Kim analysis, three single-cell reference datasets were considered: Dogga-*pf*, Howick-*pf*, and Howick-*pk*. Since the true stage proportions are unavailable for the real Kepple samples, performance was evaluated by comparing the agreement of estimated ring, trophozoite, and schizont proportions across different references.

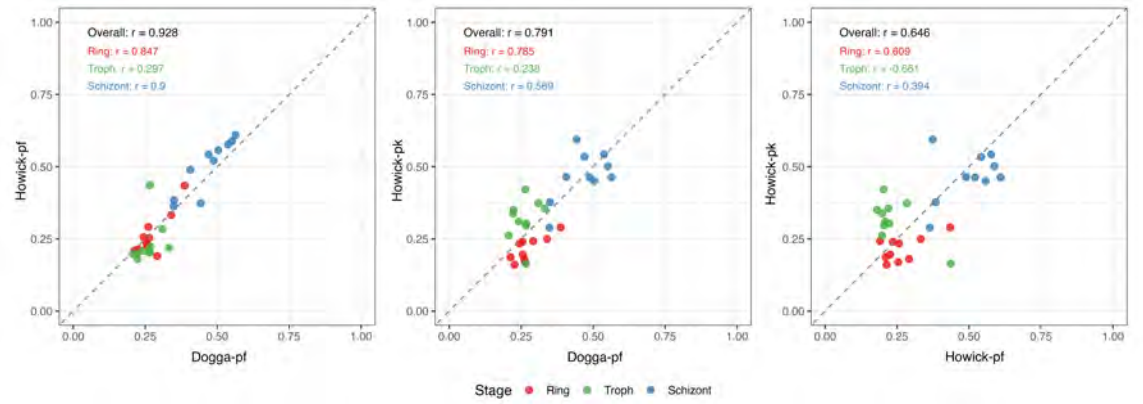
Figure 2.12 shows the pairwise comparison of estimated stage proportions across the three references. For GSNMF+, the estimates showed moderate to high agreement across references. The overall correlations were 0.928 between Dogga-*pf* and Howick-*pf*, 0.791 between Dogga-*pf* and Howick-*pk*, and 0.646 between Howick-*pf* and Howick-*pk*. Among the three stages, the schizont stage showed the strongest agreement, while the trophozoite stage showed greater reference-dependent variation. This suggests that GSNMF+ retained reasonable cross-reference consistency on the Kepple dataset, although the agreement was weaker than that observed for the mixture-Kim dataset.

BayesPrism showed very high pairwise correlations across references in this dataset. However, its estimated proportions were concentrated in narrow stage-specific ranges, with ring proportions close to zero and trophozoite proportions consistently high across samples. CIBERSORTx showed stronger reference dependence, especially when comparing estimates based on Dogga-*pf* with those based on Howick-*pf* or Howick-*pk*. This indicates that its estimated stage proportions were more sensitive to the choice of reference dataset.

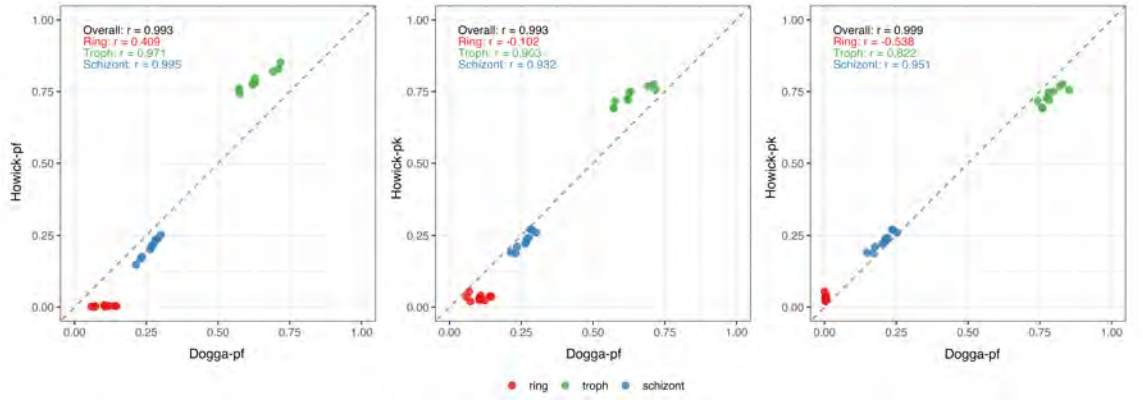
Figure 2.13 further compares the estimated stage proportions across individual Kepple samples. For GSNMF+, the three reference-specific trajectories generally followed similar sample-level patterns, especially for the schizont stage. Some separation was observed for the trophozoite stage, indicating that this stage was more sensitive

to reference choice in the Kepple dataset. BayesPrism produced smoother and more reference-consistent curves, but the estimated ring proportions remained close to zero for most samples. CIBERSORTx showed the largest disagreement across references, with substantially different stage allocations depending on the reference dataset used.

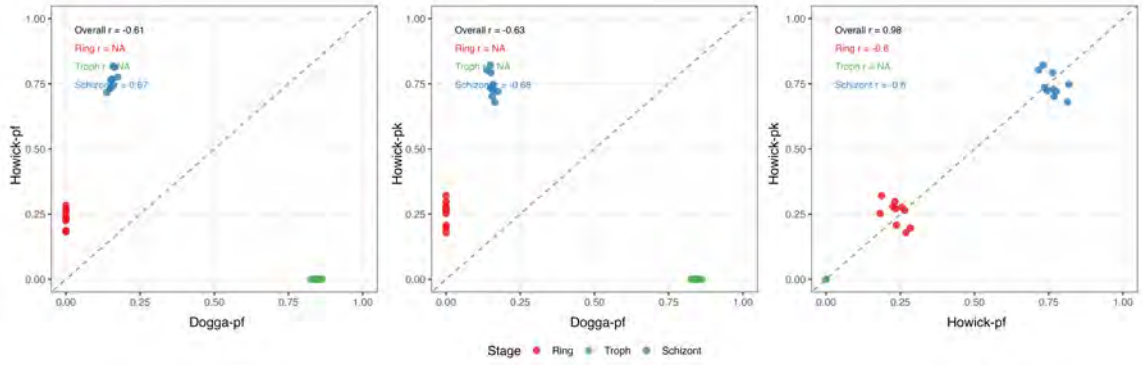
Overall, although GSNMF+ did not achieve the strongest reference consistency for the mixture-Kepple dataset, its performance remained comparable to the existing methods. In particular, GSNMF+ produced broadly consistent stage-proportion patterns across the three reference datasets, especially for the schizont stage. These results suggest that, even in a smaller and potentially more challenging real bulk dataset, GSNMF+ can still provide competitive and biologically interpretable deconvolution results. Together with the mixture-Kim analysis, the Kepple results indicate that GSNMF+ remains reasonably robust across real malaria bulk RNA-seq datasets, even when reference effects and cross-species differences are present.



(a)

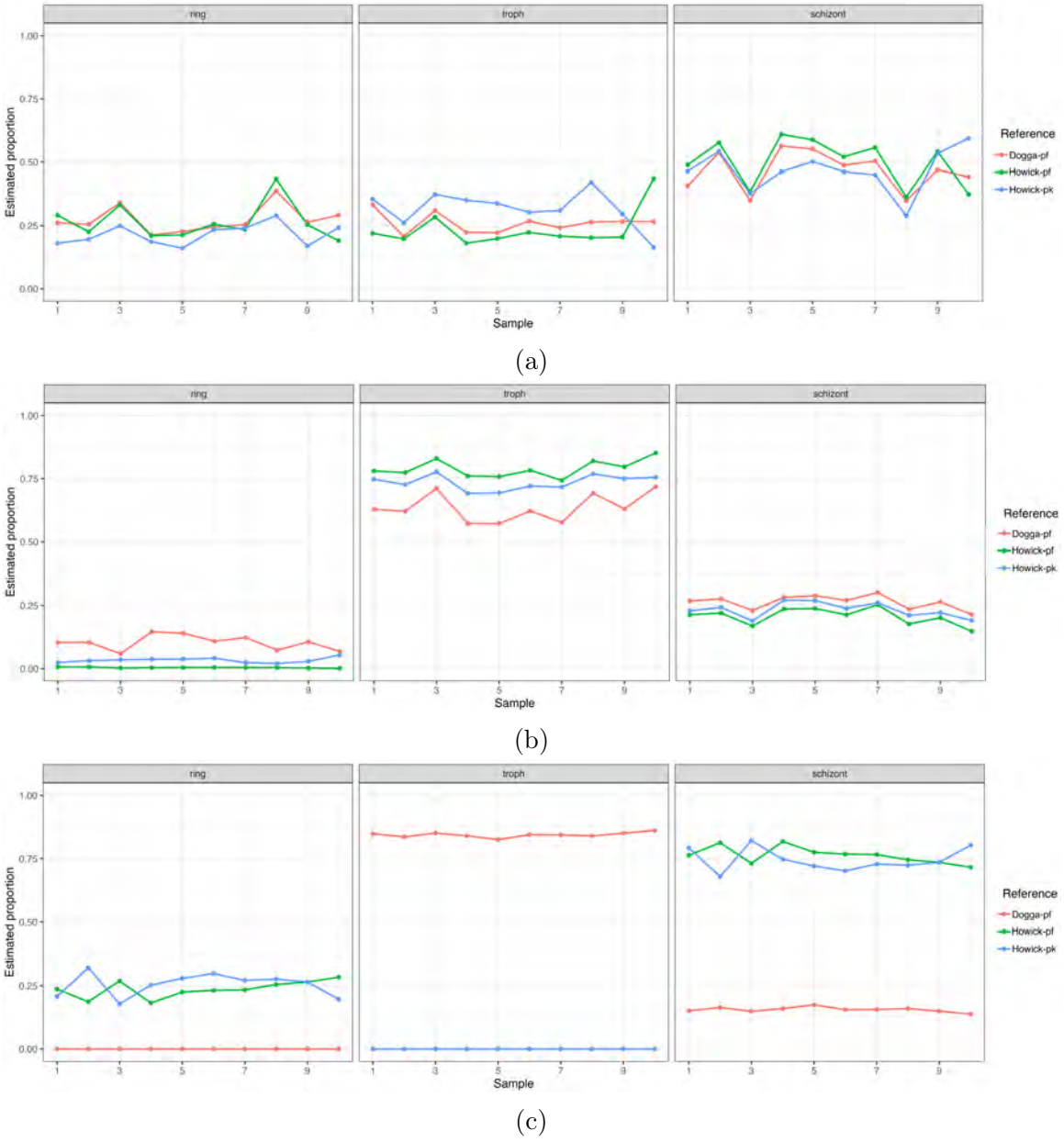


(b)



(c)

Figure 2.12: Pairwise comparison of estimated stage proportions obtained using the Dogga-pf, Howick-pf, and Howick-pk reference datasets for the mixture-Kepple dataset. Colors denote Ring, Trophozoite, and Schizont stages. The dashed line indicates perfect agreement. (a) GSNMF+. (b) BayesPrism. (c) CIBERSORTx.



(a)

(b)

(c)

Figure 2.13: Pairwise comparison of estimated stage proportions obtained using the Dogga, Howick, and PK reference datasets for the mixture-kepple dataset. Colors denote Ring, Troph, and Schizont stages. The dashed line indicates perfect agreement. (a) GSNMF+. (b) BayesPrism. (c) CIBERSORTx

2.6.3 Application to Human Pancreatic Islets

To evaluate whether GSNMF+ can be generalized beyond malaria deconvolution, we further applied it to a type 2 diabetes (T2D) pancreatic islet bulk RNA-seq dataset. In this analysis, the real bulk samples were decomposed into four endocrine cell types:

alpha, beta, delta, and gamma cells. Two independent single-cell pancreas reference datasets, Xin[63] and EMTAB[64], were used to assess the reference consistency of each method. As in the malaria real-data analysis, the true cell-type proportions are unknown for the real bulk samples; therefore, the evaluation focuses on the agreement of estimated proportions obtained using different references.

The pairwise comparison between estimates obtained by using the Xin and EMTAB references is shown in Figure 2.14. GSNMF+ achieved the strongest overall agreement between the two references, with an overall Pearson correlation of 0.975 and an overall MSE of 0.0017. This was the lowest MSE among all compared methods, indicating that GSNMF+ produced the most reference-consistent estimates in the T2D application. CIBERSORTx also showed a high overall correlation, but its MSE was larger than that of GSNMF+, and the estimates showed systematic shifts for some cell types. MuSiC showed good overall correlation but larger discrepancies for alpha and beta cells. BayesPrism obtained the weakest overall agreement among the four methods.

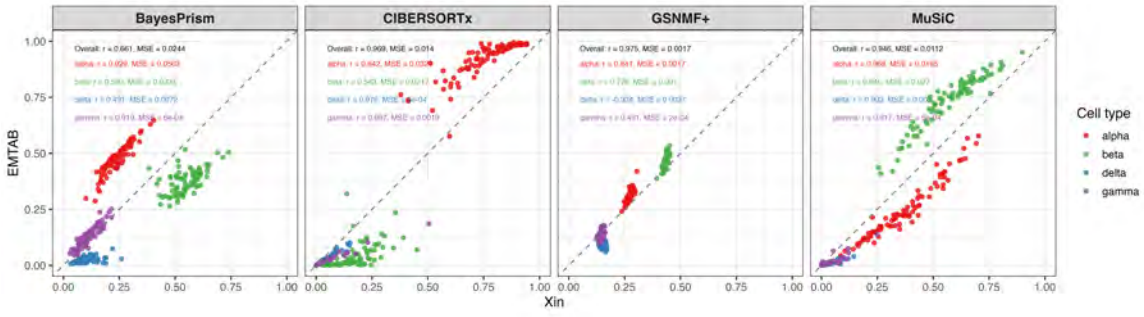


Figure 2.14: Pairwise comparison of estimated cell-type proportions obtained using the Xin and EMTAB scRNAseq reference datasets for the T2D bulk RNA-seq dataset. Each point represents one sample-cell-type estimate. The dashed line indicates perfect agreement between estimates.

Figure 2.15 shows the estimated cell-type proportions across samples using the Xin and EMTAB references. For GSNMF+, the estimates based on the two references

were highly consistent across samples and cell types. The two GSNMF+ curves were closely aligned, especially for the alpha, delta, and gamma cell types, indicating that the estimated proportions were relatively insensitive to the choice of reference dataset. In contrast, the other methods showed larger reference-dependent variation. CIBERSORTx produced markedly different estimates for several cell types, especially alpha and beta cells. MuSiC also showed larger sample-to-sample fluctuations and stronger discrepancies between the two references. BayesPrism produced smoother estimates but still showed visible separation between the Xin- and EMTAB-based results for some cell types.

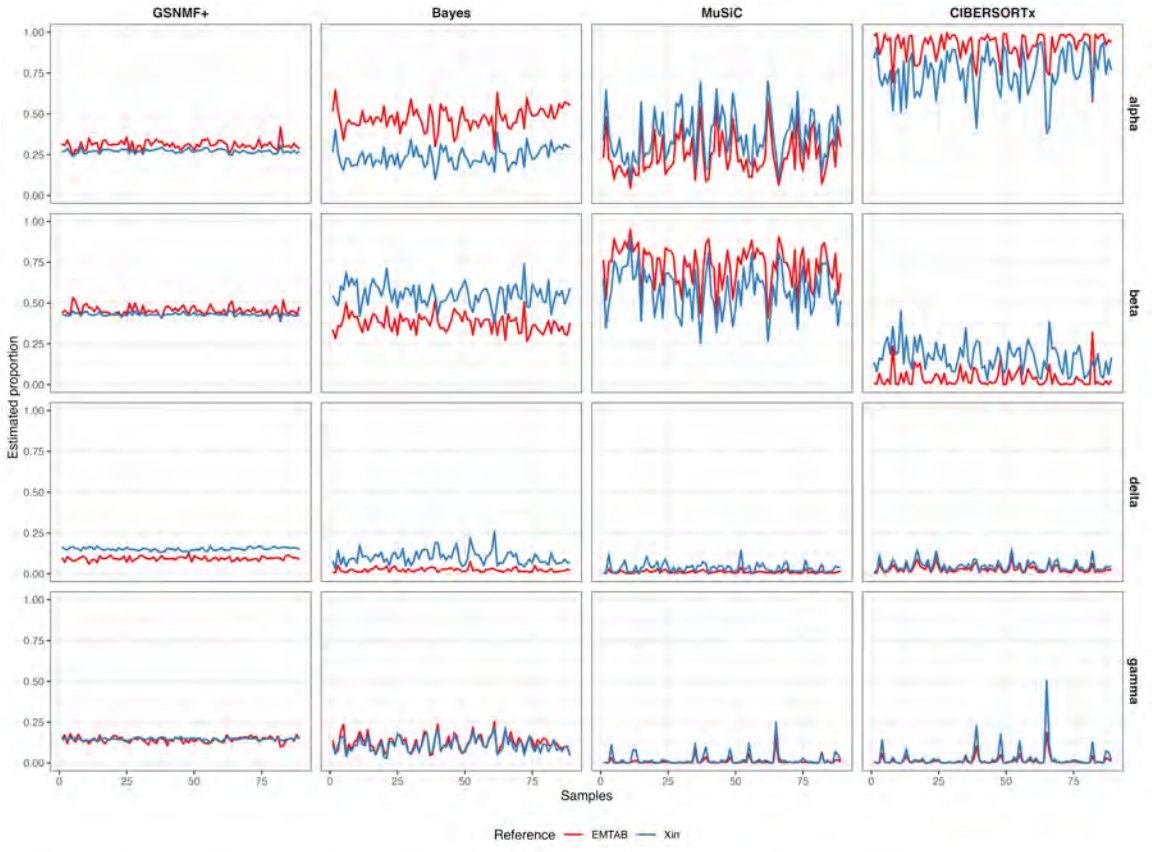


Figure 2.15: Estimated cell-type proportions across T2D bulk RNA-seq samples using the Xin and EMTAB single-cell reference datasets. Columns correspond to different deconvolution methods, and rows correspond to endocrine cell types. Within each panel, the two lines represent estimates obtained using the Xin and EMTAB references, respectively.

Overall, the T2D analysis demonstrates that GSNMF+ can be applied beyond malaria transcriptomic data and can provide stable deconvolution results in a heterogeneous human tissue setting. The strong agreement between Xin- and EMTAB-based estimates suggests that the proposed framework is robust to the use of independent single-cell references. These results support the broader applicability of GSNMF+ for bulk RNA-seq deconvolution across different biological systems.

2.7 Discussion

In this chapter, we introduce the GSNMF+, a geometry-guided nonnegative matrix factorization framework with data augmentation for robust bulk RNA-seq deconvolution. The method was developed to address several practical limitations of the original GS-NMF framework, including limited coverage of the composition space, permutation ambiguity of estimated components, and optimization complexity. By incorporating synthetically generated pseudo-bulk mixtures, GSNMF+ improves the geometric coverage of the cell-type or cell-stage proportion simplex and provides known ground-truth proportions that can be used for component annotation. In addition, the multiplicative-update-based optimization procedure simplifies implementation while preserving nonnegativity constraints.

The simulation studies demonstrated that GSNMF+ can accurately recover known stage proportions across a range of compositional settings. In particular, the method showed strong performance in stage-dominant scenarios, where the observed pseudo-bulk mixtures are concentrated near one region of the simplex and standard deconvolution methods may become unstable. The repeated simulation experiments further showed that GSNMF+ achieved lower median MSE than BayesPrism in most scenario-stage combinations, especially in the ring-dominant scenario and for ring-stage estimation. These results suggest that the augmentation strategy improves robustness when the original bulk samples alone do not provide sufficient geometric information for stable deconvolution.

The real-data applications further supported the practical utility of GSNMF+. For malaria bulk RNA-seq datasets, where true parasite stage proportions are unavailable, we evaluated performance by comparing the consistency of estimated proportions across independent single-cell reference datasets. In the mixture-Kim dataset, GSNMF+ produced more consistent estimates across Dogga-*pf*, Howick-*pf*, and Howick-*pk* references than the comparison methods, suggesting reduced sensitivity to reference choice. In the mixture-Kepple dataset, GSNMF+ did not achieve the strongest reference consistency in every comparison, but its estimates remained comparable to the existing methods and biologically interpretable. The type 2 diabetes application further showed that GSNMF+ can be generalized beyond malaria and can produce stable deconvolution results in a heterogeneous human tissue setting.

An important practical consideration in GSNMF+ is the selection of marker genes. Marker genes play a central role in the solvability constraint because they provide biological anchors for the geometric structure of the decomposition. If the selected marker genes are highly stage-specific and consistently expressed, they can help guide the estimated expression matrix toward interpretable cell-stage directions. However, if the marker genes are noisy, weakly stage-specific, or incorrectly assigned, the solvability constraint may introduce bias into the factorization. Therefore, marker gene selection is a crucial step in the framework. In practice, the number of marker genes and the criteria used to identify them should be chosen carefully based on the sparsity, noise level, and biological quality of the available reference data.

There are several limitations of the current framework. First, the performance of GSNMF+ still depends on the quality and biological relevance of the single-cell reference data used to generate augmentation samples. Although augmentation improves robustness to reference mismatch, it cannot fully remove biological differences between the reference data and the target bulk samples. In particular, different single-cell datasets may have different expression scales, sparsity levels, and stage-specific

expression patterns due to sequencing depth, experimental protocol, batch effects, or biological variation. These differences may affect the pseudo-bulk profiles generated during augmentation and therefore influence the final deconvolution results. Second, the current augmentation strategy assumes that pseudo-bulk mixtures generated from single-cell references are representative of real bulk mixtures. However, real bulk RNA-seq data may differ from aggregated single-cell profiles because of differences in library preparation, normalization, dropout, and technical noise.

Future work can further improve GSNMF+ in several directions. First, more adaptive marker gene selection strategies could be developed to reduce the influence of noisy or incorrectly assigned marker genes. Since the solvability constraint relies on marker-gene information, improving the reliability of marker gene selection may further enhance deconvolution accuracy and interpretability. Second, future extensions could integrate information from multiple references simultaneously rather than relying on one reference dataset at a time. This may further improve robustness when individual references are incomplete or biased. Finally, the framework could be evaluated on additional biological systems and larger multi-reference studies to further assess its generalizability.

CHAPTER 3: BetaDE: detecting temporal differential expression and gene programs with beta basis functions

3.1 Introduction

Single-cell RNA sequencing provides a powerful framework for studying dynamic biological processes at single-cell resolution. In many applications, including malaria parasite development, cells or parasites are observed at different stages of an underlying biological process. After cells are ordered along a trajectory using pseudotime estimation, an important downstream task is to identify genes whose expression changes systematically along pseudotime and to characterize their temporal expression patterns. These temporally dynamic genes may reveal stage-specific transcriptional programs, regulatory mechanisms, and coordinated gene modules associated with biological progression.

Existing methods for pseudotime-based gene expression analysis often model expression as a smooth function of pseudotime using splines, generalized additive models, Gaussian processes, or other flexible regression frameworks. These approaches have been useful for detecting genes associated with developmental trajectories. However, single-cell gene expression data remain challenging to analyze because they are high-dimensional, sparse, noisy, and often contain a large fraction of zero counts. In addition, gene expression counts may exhibit overdispersion and zero inflation, making simple continuous or Poisson-based models insufficient for accurately capturing the distributional features of the data.

Another important challenge is the representation of temporal expression patterns. Genes may show diverse dynamic behaviors, including early activation, late activation, transient expression, monotonic increase, monotonic decrease, or stage-specific peaks. Directly clustering raw expression profiles may obscure these patterns because of technical noise and cell-to-cell variability. Therefore, a useful temporal modeling

framework should not only identify genes with significant expression changes along pseudotime, but also provide interpretable functional features that summarize each gene's temporal behavior and support downstream clustering for gene module discovery.

To address these challenges, this chapter proposes BetaDE, a beta basis function-based framework for temporal gene expression analysis in single-cell RNA-seq data. BetaDE represents temporal expression dynamics using a collection of beta distribution-based basis functions defined on the pseudotime interval. Because beta density functions can take a wide range of shapes and peak locations, they provide a flexible and interpretable representation for capturing diverse temporal patterns. For each gene, BetaDE fits statistical models using beta-kernel features and considers multiple count distributions, including Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial models. The best-fitting distribution is selected using the Akaike Information Criterion (AIC), allowing the method to adapt to different levels of dispersion and zero inflation across genes.

Beyond detecting temporally differentially expressed genes, BetaDE uses the estimated beta-kernel coefficients as functional features to summarize gene-level temporal expression patterns. These coefficient-based representations reduce the influence of noisy raw expression values and provide low-dimensional features for clustering genes into temporal modules. The resulting modules can then be interpreted biologically through their temporal patterns and functional annotations.

This chapter evaluates BetaDE using both simulation studies and real single-cell malaria transcriptomic datasets. In the simulation studies, we assess the ability of BetaDE to detect temporally dynamic genes and recover meaningful temporal patterns under different count distributions and noise settings. In the real data analysis, we apply BetaDE to single-cell malaria datasets to identify dynamic genes and gene modules associated with parasite developmental stages, including ring, trophozoite,

and schizont stages. Through these analyses, BetaDE provides a unified framework for temporal differential expression analysis, functional feature extraction, and gene module discovery in single-cell temporal transcriptomic data.

The remainder of this chapter is organized as follows. Section 3.2 presents the statistical framework of BetaDE, including beta basis function construction, model specification, model selection, and inference procedures. Section 3.3 describes the single-cell datasets used in this project. Section 3.4 presents the simulation design and simulation results used to evaluate the performance of the proposed method. Section 3.5 validates the proposed framework using data with known time-point information. Section 3.6 applies BetaDE to real single-cell malaria transcriptomic datasets and presents the model fitting and clustering results. Finally, Section 3.7 summarizes the main findings, discusses limitations, and outlines possible extensions.

3.2 BetaDE Method

This section presents the proposed BetaDE framework for temporal differential expression analysis in single-cell RNA-seq data. Given a single-cell expression count matrix and an inferred pseudotime vector, BetaDE models gene expression as a function of pseudotime using a library of beta basis functions. The method is designed to identify genes whose expression changes significantly along pseudotime and to provide interpretable temporal features for downstream gene clustering.

3.2.1 Overview of the BetaDE framework

BetaDE takes as input a single-cell RNA-seq count matrix and a pseudotime vector estimated from a trajectory inference method. Let $Y = (Y_{ij}) \in \mathbb{N}^{n \times m}$ denote the expression count matrix, where Y_{ij} is the observed count of gene j in cell i, $i = 1, \ldots, n$, and $j = 1, \ldots, m$. Let $t_i \in (0, 1)$ denote the normalized pseudotime value of cell i. If pseudotime values are exactly 0 or 1, a small numerical adjustment is applied to avoid boundary issues when evaluating beta density functions.

The rational of BetaDE is to represent possible temporal expression patterns using beta basis functions. Each beta kernel corresponds to a candidate temporal pattern, such as early activation, late activation, transient expression, or stage-specific peak expression. For each gene, BetaDE fits count-based regression models using these beta-kernel features and evaluates whether gene expression is significantly associated with pseudotime. The BetaDE workflow consists of five main steps. First, a library of beta kernel basis functions is constructed on the pseudotime interval. Second, each beta kernel is evaluated at the pseudotime value of every cell to obtain temporal kernel features. Third, for each gene-kernel pair, multiple count models are fitted, including Poisson, Negative Binomial, zero-inflated Poisson, and zero-inflated Negative Binomial models. Fourth, model selection is performed based on the Akaike Information Criterion (AIC). Finally, significant genes are clustered based on their fitted beta-kernel coefficients to identify temporal gene modules.

3.2.2 Beta kernel basis construction

To flexibly represent diverse temporal gene expression patterns, BetaDE constructs a library of K beta kernel basis functions defined on the interval $(0, 1)$. For the k-th beta kernel, let $a_k > 0$ and $b_k > 0$ denote the two shape parameters. The beta kernel function is defined as

$$f_k(t) = \text{Beta}(t; a_k, b_k) = \frac{t^{a_k-1}(1-t)^{b_k-1}}{B(a_k, b_k)}, \qquad 0 < t < 1, \tag{3.1}$$

where $B(a_k, b_k) = \Gamma(a_k)\Gamma(b_k)/\Gamma(a_k + b_k)$ denotes the beta function. For each cell i and beta kernel k, the kernel function is evaluated at the normalized pseudotime value t_i to define the temporal kernel feature $\phi_{ik} = f_k(t_i)$, for $i = 1, \ldots, n$ and $k = 1, \ldots, K$. The collection of these features forms the beta-kernel feature matrix $\Phi = (\phi_{ik}) \in \mathbb{R}^{n\times K}$.

The beta density family is particularly useful for pseudotime analysis because it can

generate a wide range of shapes on a bounded interval. Depending on the values of (a_k, b_k), beta kernels can represent left-skewed patterns corresponding to early activation, right-skewed patterns corresponding to late activation, symmetric unimodal patterns corresponding to transient expression, and nearly monotonic temporal trends. Therefore, the beta kernel library serves as an interpretable dictionary of candidate gene expression dynamics along pseudotime. In this study, as shown in Figure 3.1, we use $K = 22$ beta kernel basis functions to cover a broad range of temporal expression shapes. These kernels are fixed before model fitting and are applied to all genes.

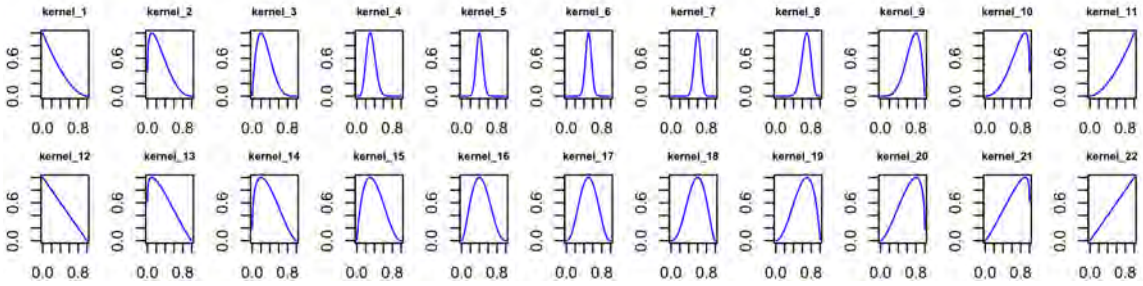


Figure 3.1: 22 Beta Basis Functions with Varying Shapes and Locations

3.2.3 Statistical model for gene expression

For each gene j, BetaDE models the observed expression count Y_{ij} as a function of a beta-kernel feature ϕ_{ik}. Instead of fitting all beta kernels jointly in a single regression model, BetaDE adopts a one-kernel-at-a-time strategy. Under this strategy, each beta kernel represents one candidate temporal expression pattern, and the association between gene expression and this temporal pattern is evaluated separately. For each gene-kernel pair (j, k), BetaDE considers four candidate read count models: Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial models. These models are fitted using the R package `glmmTMB`, which provides a unified framework for fitting count and zero-inflated regression models by maximum likelihood. In the implementation of BetaDE, the beta-kernel feature ϕ_{ik} is included as the predictor in the conditional mean model, and the candidate distribution family

is specified through the corresponding `glmmTMB` model family.

3.2.3.1 Poisson model

For each gene j and beta kernel k, BetaDE first fits a Poisson regression model to evaluate whether the expression of gene j is associated with the temporal pattern represented by kernel k. Specifically, the model is defined as

$$\begin{cases} Y_{ij} \sim \text{Poisson}(\lambda_{jk}), \\ \log((\lambda_{jk}|t_i)) = \beta_{0,jk} + \beta_{1,jk}\phi_{ik}. \end{cases} \tag{3.2}$$

Here, λ_{jk} denotes the expected expression level of gene j under kernel k. The parameter $\beta_{0,jk}$ represents the baseline expression level, while $\beta_{1,jk}$ measures the association between gene expression and the beta-kernel feature. A significant nonzero value of $\beta_{1,jk}$ suggests that the expression of gene j changes along pseudotime according to the temporal pattern captured by kernel k.

3.2.3.2 Negative binomial model

Single-cell RNA-seq counts often show overdispersion, where the variance is larger than the mean. To account for this property, BetaDE also considers the negative binomial model. For each gene j and beta kernel k, the model is defined as

$$\begin{cases} Y_{ij} \sim \text{NB}(\mu_{jk}, \theta_{jk}), \\ \log((\mu_{jk}|t_i)) = \beta_{0,jk} + \beta_{1,jk}\phi_{ik}. \end{cases} \tag{3.3}$$

Here, μ_{jk} denotes the expected expression level of gene j and θ_{jk} is the dispersion parameter. The parameter $\beta_{0,jk}$ represents the baseline expression level of gene j, while $\beta_{1,jk}$ measures the association between gene expression and the beta-kernel feature ϕ_{ik}. Compared with the Poisson model, the negative binomial model provides additional flexibility by allowing the variance to exceed the mean.

3.2.3.3 Zero-inflated Poisson model

To account for excess zeros, BetaDE further considers a zero-inflated Poisson model. Let Z_{ijk} be a latent indicator for whether the observation belongs to the Poisson count component, and let p_{ijk} denote the probability that the observation belongs to this component. For each gene j and beta kernel k, the model is defined as

$$\begin{cases} Z_{ijk} \sim \text{Ber}(p_{ijk}), \\ Y_{ij} \mid Z_{ijk} \sim Z_{ijk} \cdot \text{Poisson}(\lambda_{jk}) + (1 - Z_{ijk}) \cdot 0, \\ \log((\lambda_{jk}|t_i)) = \beta_{0,jk} + \beta_{1,jk}\phi_{ik}, \\ \text{logit}(p_{ijk}) = \gamma_{0,jk} + \gamma_{1,jk} \log(\lambda_{jk}). \end{cases} \tag{3.4}$$

Here, λ_{jk} denotes the mean of the Poisson count component, and p_{ijk} denotes the probability of belonging to the Poisson count component. Therefore, $1 - p_{ijk}$ represents the probability of belonging to the structural zero component. The parameters $\beta_{0,jk}$ and $\beta_{1,jk}$ model the temporal change in the count component, while $\gamma_{0,jk}$ and $\gamma_{1,jk}$ model how the probability of belonging to the count component varies with the expected expression level.

3.2.3.4 Zero-inflated negative binomial model

Finally, BetaDE considers a zero-inflated negative binomial model to account for both overdispersion and excess zeros in single-cell RNA-seq count data. Let Z_{ijk} be a latent indicator for whether the observation belongs to the negative binomial count component, and let p_{ijk} denote the probability that the observation belongs to this component. For each gene j and beta kernel k, the model is defined as

$$\begin{cases} Z_{ijk} \sim \text{Ber}(p_{ijk}), \\ Y_j \mid Z_{ijk} \sim Z_{ijk} \cdot \text{NB}(\mu_{jk}, \theta_{jk}) + (1 - Z_{ijk}) \cdot 0, \\ \log((\mu_{jk}|t_i)) = \beta_{0,jk} + \beta_{1,jk}\phi_{ik}, \\ \text{logit}(p_{ijk}) = \gamma_{0,jk} + \gamma_{1,jk} \log(\mu_{jk}). \end{cases} \tag{3.5}$$

Here, μ_{jk} denotes the mean of the negative binomial count component, and θ_{jk} is the dispersion parameter. The probability p_{ijk} represents the probability of belonging to the count component, while $1 - p_{ijk}$ represents the probability of belonging to the structural zero component. Compared with the zero-inflated Poisson model, the zero-inflated negative binomial model provides additional flexibility by accounting for overdispersion in the count component.

3.2.4 Model selection

For each gene-kernel pair (j, k), the four candidate models are fitted independently using the R package `glmmTMB`. Specifically, the Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial models are fitted by specifying the corresponding conditional distribution and zero-inflation structure in `glmmTMB`. Models that fail to converge or produce invalid parameter estimates are excluded from further comparison. For each successfully fitted model, BetaDE computes the Akaike Information Criterion, defined as $\text{AIC} = -2\ell + 2p$, where ℓ is the maximized log-likelihood returned by the fitted `glmmTMB` model and p is the number of estimated parameters.

Because zero-inflated models introduce additional parameters, BetaDE uses a conservative AIC rule before allowing a zero-inflated model to enter the final model comparison. The Poisson and negative binomial models are retained as non-zero-inflated baseline models. The zero-inflated Poisson model is retained only if its AIC is at least 10 units smaller than that of the corresponding Poisson model, $\text{AIC}_{jk,\text{ZIP}} \leq \text{AIC}_{jk,\text{Poisson}} - 10$. Similarly, the zero-inflated negative binomial model is retained only if its AIC is at least 10 units smaller than that of the corresponding negative binomial model, $\text{AIC}_{jk,\text{ZINB}} \leq \text{AIC}_{jk,\text{NB}} - 10$. This threshold follows the commonly used interpretation of AIC differences, where an AIC difference greater than 10 indicates substantially weaker empirical support for the model with the larger AIC [68].

After applying this filtering rule, the best-fitting distributional model for each gene-kernel pair is selected from the retained candidate models as $d^*_{jk} = \arg\min_{d \in \mathcal{D}_{jk}} \mathrm{AIC}_{jkd}$, where $\mathcal{D}_{jk}$ denotes the set of successfully fitted and retained candidate models for gene j and kernel k. After selecting the best distributional model for each gene-kernel pair, BetaDE further identifies the best-supported temporal kernel for each gene by comparing the selected AIC values across all beta kernels, $k^*_j = \arg\min_k \mathrm{AIC}_{jkd^*_{jk}}$. The selected pair $(k^*_j, d^*_{jk^*_j})$ represents the temporal kernel shape and count distribution that best explain the observed expression pattern of gene j. This model selection procedure allows BetaDE to adapt to both temporal heterogeneity and distributional heterogeneity across genes, while avoiding unnecessary zero-inflation unless it provides a substantial improvement in model fit.

3.2.5 Differential expression detection

After model selection, BetaDE tests whether each gene shows significant temporal variation along pseudotime. For gene j, let k^*_j denote the selected beta kernel. The fitted mean model can be written as $\log(\mu_j(t_i)) = \hat{\beta}_{0,j} + \hat{\beta}_{1,j}\phi_{ik^*_j}$, where μ_{ij} is the expected expression level of gene j at pseudotime t_i, and $\phi_{ik^*_j}$ is the selected beta-kernel feature evaluated at the pseudotime.

For each gene, the significance of $\beta_{1,j}$ is evaluated from the fitted `glmmTMB` model, and the corresponding p-value is used to assess whether gene expression is significantly associated with the selected beta-kernel feature. The resulting p-values are adjusted across genes using the Benjamini–Hochberg procedure to control the false discovery rate. Genes with adjusted p-values below the significance threshold are identified as temporally differentially expressed genes.

3.2.6 Functional clustering of significant genes

After identifying temporally differentially expressed genes, BetaDE further clusters significant genes based on coefficient-derived functional features. The goal of this

step is to group genes with similar pseudotime-dependent expression patterns into temporal gene modules.

For each gene, we constructed a gene-by-feature matrix for clustering. For the beta kernels $k = 1, \ldots, K$, the feature values were defined as the estimated kernel effect coefficients $\widehat{\beta}_{1,jk}$, where $\widehat{\beta}_{1,jk}$ measures the association between the expression of gene j and beta kernel k. If a constant kernel is included in the kernel library, its corresponding feature is represented by the fitted intercept coefficient. Therefore, each gene is summarized by its fitted coefficients across the kernel library.

In addition to the coefficient-based features, three gene-level summary features were included: the significance indicator from the selected best-fitting kernel model, together with the location and variance of the selected kernel. The significance indicator records whether the gene is identified as temporally significant under its selected model, while the location and variance features summarize the center and spread of the selected beta kernel along pseudotime.

Let $\mathbf{x}_j$ denote the final feature vector for gene j. The collection of feature vectors forms a gene-by-feature matrix $X_{\mathcal{G}} = (\mathbf{x}_j)_{j \in \mathcal{G}}$, where $\mathcal{G}$ is the set of significant DE genes after multiple testing correction. Clustering is then applied to $X_{\mathcal{G}}$ to group genes with similar coefficient patterns and temporal characteristics. In practice, different clustering algorithms can be used on this feature matrix, such as hierarchical clustering, k-means clustering, or other distance-based clustering methods. The choice of clustering method depends on the desired cluster structure and the downstream interpretation.

The resulting clusters are interpreted as temporal gene modules. Genes within the same cluster are expected to share similar pseudotime-dependent expression behavior, such as early activation, intermediate peak expression, or late-stage activation. These modules provide a basis for downstream biological interpretation and functional enrichment analysis.

3.3 Data Description

Table 3.1: Information about the three malaria single-cell RNA sequencing datasets used in this study.

Dataset	**Source**	**Strain**	**No. of genes**	**No. of cells**	**Cells at stage**
Howick-pf [13]	Laboratory	Pf3D7	3,595	6,493	R:1,251, T:4,477, S:765
Dogga-pf [58]	Laboratory	Pf3D7	4,923	29,216	R:7,499, T:13,558, S:8,159
Painter-pf [69]	Laboratory	Pf3D7	5,198	Time points: 0–47 hpi	

R = Ring; T = Trophozoite; S = Schizont. hpi = hours post-invasion.

This study uses three single-cell RNA sequencing datasets of *Plasmodium falciparum* generated from laboratory Pf3D7 parasites. These datasets provide complementary information for studying transcriptional dynamics during the intraerythrocytic developmental cycle. The Howick-pf and Dogga-pf datasets contain single-cell expression profiles with cells annotated by developmental stage, including ring, trophozoite, and schizont stages. The Painter-pf dataset contains time-resolved single-cell transcriptomic profiles collected across 0–47 hours post invasion (hpi), providing known experimental time-point information for validating pseudotime-based temporal patterns.

Table 3.1 summarizes the datasets used in this chapter. The Howick-pf [13] dataset contains 3,595 genes and 6,493 cells, including 1,251 ring-stage cells, 4,477 trophozoite-stage cells, and 765 schizont-stage cells. The Dogga-pf [58] dataset contains 4,923 genes and 29,216 cells, including 7,499 ring-stage cells, 13,558 trophozoite-stage cells, and 8,159 schizont-stage cells. Compared with the Howick-pf dataset, the Dogga-pf dataset includes a larger number of cells and provides broader coverage of the parasite developmental stages. The Painter-pf [69] dataset contains 5,198 genes and single-cell profiles collected from 0 to 47 hpi, which enables comparison between inferred pseudotime and experimentally measured developmental time.

Before applying BetaDE, genes were filtered and matched across datasets as required for each analysis. Specifically, genes were retained only if they were expressed in at least 100 cells in both the Howick-pf and Dogga-pf datasets. Pseudotime values were inferred from the single-cell RNA sequencing data using trajectory inference and subsequently rescaled to the interval (0,1). The stage annotations available for the Howick-pf and Dogga-pf datasets were used to evaluate whether the temporal gene modules identified by BetaDE were consistent with known developmental stages of Plasmodium falciparum. In addition, the experimentally measured time points in the Painter-pf dataset served as an external reference for assessing whether the inferred temporal expression patterns agreed with the known progression of the parasite life cycle.

3.4 Simulation

3.4.1 Simulation setting

To evaluate the performance of BetaDE in recovering both the underlying count distribution and the kernel-specific temporal expression pattern, we conducted simulation studies using a synthetic single-cell RNA-seq count matrix. The simulated dataset contained 500 genes measured across 1,000 cells. For each cell i, the pseudotime value t_i was independently sampled from a uniform distribution on $(0, 1)$. To generate gene-specific temporal patterns, we used five predefined kernel groups derived from the beta kernel library: Group 1: $(1, 12)$, Group 2: $(3, 14)$, Group 3: $(6, 17)$, Group 4: $(8, 19)$, and Group 5: $(10, 21)$.

These groups were chosen to represent diverse temporal expression patterns, including early, intermediate, and late activation along pseudotime. The 500 genes were divided equally into four data-generating model classes: Poisson, negative binomial (NB), zero-inflated Poisson (ZIP), and zero-inflated negative binomial (ZINB), with 125 genes assigned to each class. Therefore, 50% of genes were generated from zero-inflated models. Within each model class, genes were distributed as evenly as

possible across the five kernel groups, and each gene was assigned one true kernel from its corresponding group.

For gene j, let k_j denote the assigned true kernel. For cell i, the corresponding kernel feature was defined as $\phi_{ik_j} = f_{k_j}(t_i)$, where $f_{k_j}(\cdot)$ is the beta kernel function evaluated at pseudotime t_i. The mean expression level was generated according to $\log(\mu_{ij}) = \beta_{0j} + \beta_{1j}\phi_{ik_j}$, where β_{0j} represents the gene-specific baseline expression level and β_{1j} controls the strength of the pseudotime-associated signal. Two simulation designs were considered. In Design 1, the baseline parameter was fixed at $\beta_{0j} = 0$, and the temporal effect size was sampled from $\beta_{1j} \sim \text{Uniform}(1, 3)$. In Design 2, the baseline parameter was sampled from $\beta_{0j} \sim \text{Uniform}(-3, 3)$, while the temporal effect size was again sampled from $\beta_{1j} \sim \text{Uniform}(1, 3)$.

Counts were then generated according to the assigned model class. For Poisson genes, $Y_{ij} \sim \text{Poisson}(\mu_{ij})$. For NB genes, counts were generated from a negative binomial distribution with mean μ_{ij} and variance $1.5\mu_{ij}$. For ZIP genes, counts were generated from a zero-inflated Poisson model with the structural zero probability fixed at 0.2. For ZINB genes, counts were generated from a zero-inflated negative binomial model with the structural zero probability fixed at 0.2 and the count-component variance equal to $1.5\mu_{ij}$(Table 3.2).

After generating the simulated count matrix, BetaDE was applied using the same model-fitting procedure described in Section 3.2. For each gene, BetaDE fitted candidate Poisson, NB, ZIP, and ZINB models across the beta kernel library. The best-fitting model was selected using AIC, and temporal differential expression was assessed based on the fitted kernel-effect coefficient. The simulation study was designed to evaluate two aspects of BetaDE performance. First, we assessed whether BetaDE could correctly recover the underlying count distribution used to generate each gene. Second, we evaluated whether the selected beta kernel or kernel group was consistent with the true temporal pattern used for data generation. These evaluations

demonstrate whether BetaDE can simultaneously adapt to heterogeneous count distributions while accurately identifying meaningful temporal expression patterns along pseudotime.

Table 3.2: Summary of the simulation designs used to evaluate BetaDE.

Design	**Genes**	**Cells**	**Models**	**Zero inflation**	β_0	β_1
Design 1	500	1000	Poisson, NB	20%	0	$U(1,3)$
Design 2			ZIP, ZINB		$U(-3,3)$	

Note: Zero inflation was introduced only for the zero-inflated models, ZIP and ZINB. Both simulation designs used five kernel groups derived from the beta kernel library: Group 1 $(1,12)$, Group 2 $(3,14)$, Group 3 $(6,17)$, Group 4 $(8,19)$, and Group 5 $(10,21)$.

3.4.2 Simulation results

3.4.2.1 Model selection performance

Under Simulation Design 1, the baseline expression parameter was fixed at $\beta_0 = 0$, and the temporal effect size was sampled from $\beta_1 \sim U(1,3)$. The performance of BetaDE under this setting is summarized in Figure 3.2. Overall, BetaDE achieved strong performance in both distributional model selection and temporal kernel recovery. The overall model selection accuracy was 0.964, the kernel selection accuracy was 1.000, and the joint recovery accuracy was 0.964.

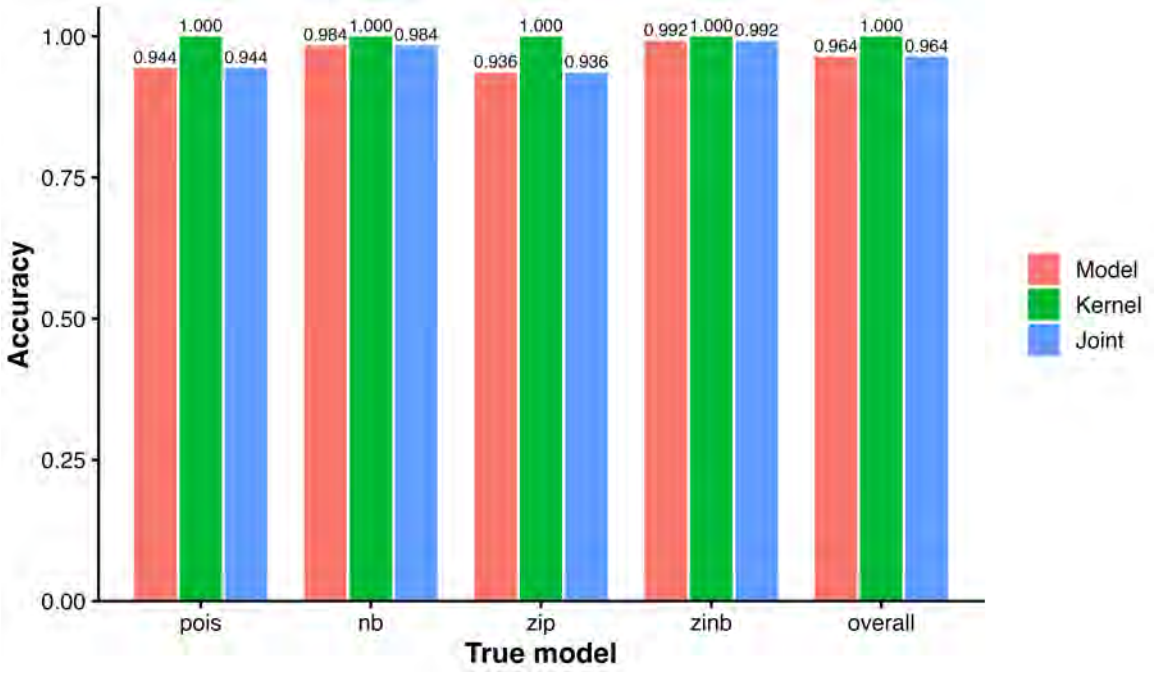


Figure 3.2: Model selection performance under Simulation Design 1

For distributional model selection, BetaDE correctly identified 118 out of 125 Poisson genes, 123 out of 125 negative binomial genes, 117 out of 125 zero-inflated Poisson genes, and 124 out of 125 zero-inflated negative binomial genes. This corresponds to model selection accuracies of 0.944, 0.984, 0.936, and 0.992 for the Poisson, NB, ZIP, and ZINB models, respectively. The overall model selection accuracy was 0.964. The main model selection errors occurred between closely related model families. For example, 7 Poisson genes were selected as NB, which is expected because the negative binomial model can approximate the Poisson model when overdispersion is weak. Similarly, 8 ZIP genes were selected as ZINB, indicating that the distinction between ZIP and ZINB models can be challenging when the additional overdispersion component is weak.

Temporal kernel recovery was perfect under Design 1. Across all 500 simulated genes, BetaDE selected the correct beta kernel for every gene, giving a kernel selection accuracy of 1.000. This indicates that the beta kernel library was able to accurately recover the underlying temporal expression patterns when the baseline expression

level was fixed and the temporal signal was moderately strong.

Because the temporal kernel was correctly recovered for all genes, the joint recovery accuracy was driven entirely by the distributional model selection accuracy. The joint accuracy was 0.944 for Poisson genes, 0.984 for NB genes, 0.936 for ZIP genes, and 0.992 for ZINB genes, with an overall joint accuracy of 0.964. These results suggest that BetaDE can accurately identify both the temporal expression pattern and the appropriate count distribution under Design 1, while the remaining errors mainly reflect the intrinsic similarity between Poisson and negative binomial models and between ZIP and ZINB models.

Table 3.3: Confusion matrix for distributional model selection under Simulation Design 1. Rows represent the true model, and columns represent the selected model.

True model	**Poisson**	**NB**	**ZIP**	**ZINB**
NB	123	0	0	2
Poisson	7	118	0	0
ZINB	0	0	124	1
ZIP	0	0	8	117

Table 3.4: Confusion matrix for temporal kernel selection under Simulation Design 1. Rows represent the true kernel, and columns represent the selected kernel.

True Kernel	**1**	**3**	**6**	**8**	**10**	**12**	**14**	**17**	**19**	**21**
1	52	0	0	0	0	0	0	0	0	0
12	0	48	0	0	0	0	0	0	0	0
3	0	0	52	0	0	0	0	0	0	0
14	0	0	0	48	0	0	0	0	0	0
6	0	0	0	0	52	0	0	0	0	0
17	0	0	0	0	0	48	0	0	0	0
8	0	0	0	0	0	0	52	0	0	0
19	0	0	0	0	0	0	0	48	0	0
10	0	0	0	0	0	0	0	0	52	0
21	0	0	0	0	0	0	0	0	0	48

Under Simulation Design 2, the baseline expression parameter was sampled from $\beta_0 \sim U(-3, 3)$, while the temporal effect size was generated from $\beta_1 \sim U(1, 3)$. Compared with Design 1, this setting introduces greater heterogeneity in baseline gene expression levels and therefore provides a more challenging scenario for model selection. The performance of BetaDE under Design 2 is summarized in Figure 3.3.

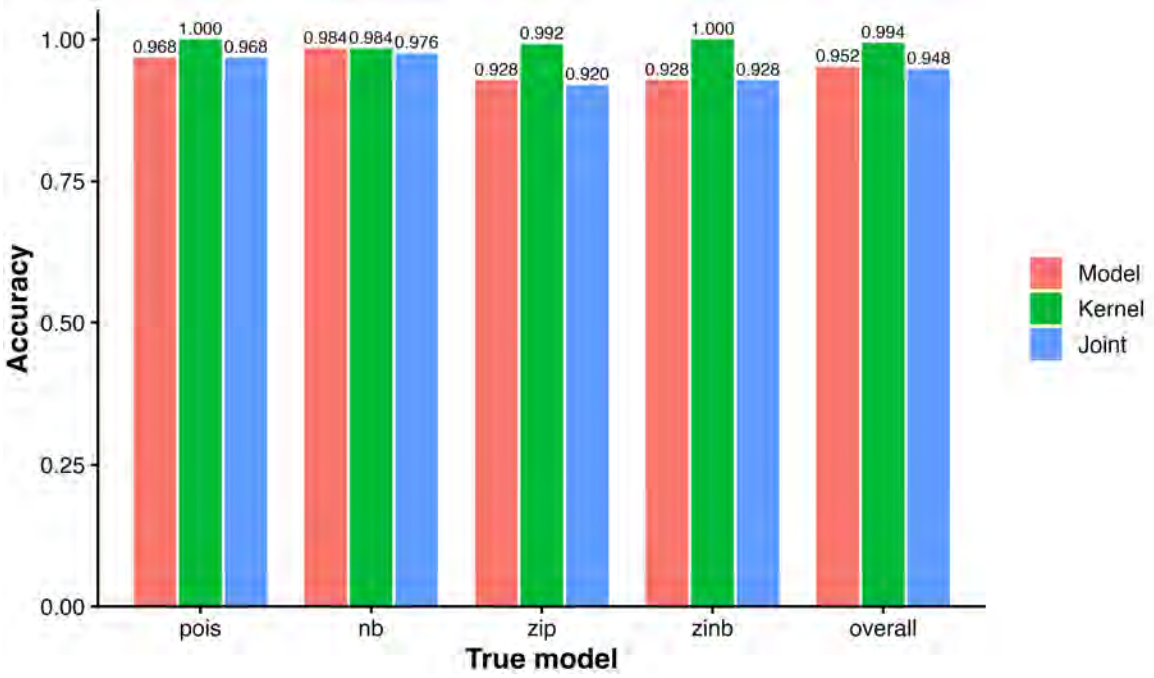


Figure 3.3: Model selection performance under Simulation Design 2

Overall, BetaDE maintained strong performance under this more heterogeneous setting. The overall distributional model selection accuracy was 0.952, the overall temporal kernel selection accuracy was 0.994, and the overall joint recovery accuracy was 0.948. For distributional model selection, BetaDE correctly identified 121 out of 125 Poisson genes, 123 out of 125 negative binomial genes, 116 out of 125 zero-inflated Poisson genes, and 116 out of 125 zero-inflated negative binomial genes. This corresponds to model selection accuracies of 0.968, 0.984, 0.928, and 0.928 for the Poisson, NB, ZIP, and ZINB models, respectively.

Most model selection errors occurred between closely related distributional families. For example, 4 Poisson genes were selected as NB. Similarly, 6 ZIP genes were selected as ZINB, and 6 ZINB genes were selected as NB, suggesting that distinguishing zero inflation from overdispersion can be challenging when baseline expression levels vary substantially across genes.

Temporal kernel recovery remained highly accurate under Design 2. Across all 500 simulated genes, BetaDE correctly selected the true temporal kernel for 497 genes,

giving an overall kernel selection accuracy of 0.994. Most kernel groups were recovered perfectly, with only three genes assigned to related kernels.

The joint recovery accuracy closely followed the distributional model selection accuracy because the kernel selection accuracy was nearly perfect. The joint accuracy was 0.968 for Poisson genes, 0.976 for NB genes, 0.920 for ZIP genes, and 0.928 for ZINB genes, with an overall joint accuracy of 0.948.

Table 3.5: Confusion matrix for distributional model selection under Simulation Design 2. Rows represent the true model, and columns represent the selected model.

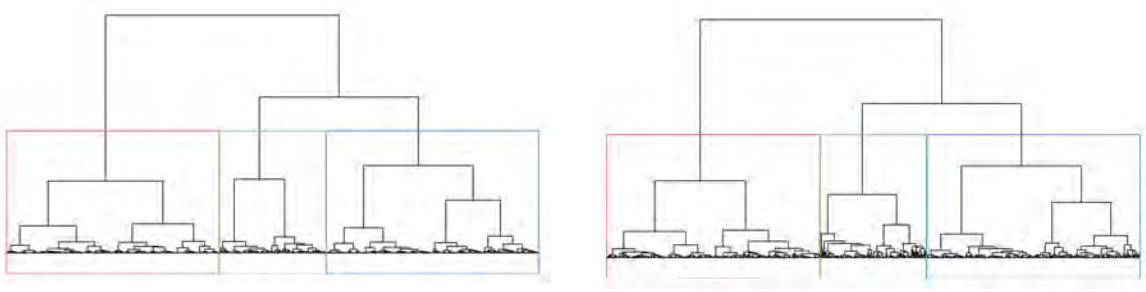
True model	**Poisson**	**NB**	**ZIP**	**ZINB**
NB	123	1	1	0
Poisson	4	121	0	0
ZINB	6	0	116	3
ZIP	3	0	6	116

Table 3.6: Confusion matrix for temporal kernel selection under Simulation Design 2. Rows represent the true kernel, and columns represent the selected kernel.

True kernel	**1**	**12**	**3**	**14**	**6**	**17**	**8**	**19**	**10**	**21**
1	51	1	0	0	0	0	0	0	0	0
12	0	48	0	0	0	0	0	0	0	0
3	0	0	52	0	0	0	0	0	0	0
14	0	0	1	47	0	0	0	0	0	0
6	0	0	0	0	52	0	0	0	0	0
17	0	0	0	0	0	48	0	0	0	0
8	0	0	0	0	0	0	52	0	0	0
19	0	0	0	0	0	0	0	48	0	0
10	0	0	0	0	0	0	0	0	52	0
21	0	0	0	0	0	0	0	0	1	47

3.4.2.2 Functional clustering based on fitted kernel features

After evaluating the accuracy of model and kernel selection, we further assessed whether the fitted kernel-based features could recover the underlying temporal expression patterns at the gene-clustering level. For each gene, a feature vector was constructed using the fitted coefficients across the 22 beta kernels. In addition, three selected-kernel summary features were included: a binary significance indicator for the selected kernel, the location of the selected kernel, and the variance of the selected kernel. Therefore, each gene was represented by a 25-dimensional feature vector. The resulting gene-by-feature matrix was row-standardized and clustered using hierarchical clustering with Ward's linkage and Euclidean distance.



(a) $\beta_0 = 0$ **(b)** $\beta_0 \sim U(-3, 3)$

Figure 3.4: First-level hierarchical clustering of simulated genes under two baseline-expression designs.

As shown in the first-level dendrograms for both simulation designs (Figure 3.4), the genes showed three major branches rather than five clearly separated branches corresponding exactly to the five true kernel groups. Therefore, the first-level clustering was interpreted as capturing broad temporal structures in the fitted kernel-based feature space. To further resolve finer temporal patterns within these broad branches, we performed second-level hierarchical clustering separately within each first-level cluster.

For the second-level clustering, genes within each first-level cluster were extracted

from the row-standardized feature matrix. If a first-level cluster contained fewer than 20 genes, it was not further divided. Otherwise, Euclidean distances were recomputed within that cluster, and hierarchical clustering was performed using Ward's linkage. Candidate numbers of subclusters were considered from $k = 2$ up to a maximum of 10, with the additional constraint that each resulting subcluster must contain at least 10 genes. For each valid candidate value of k, the average silhouette width was calculated. The candidate k that produced the highest average silhouette width was selected as the optimal number of second-level subclusters within that first-level cluster. A first-level cluster was split only if this best average silhouette width exceeded 0.15; otherwise, the cluster was retained as a single group. The final subcluster labels were constructed by combining the first-level cluster label and the second-level subcluster label, such as 1_1, 1_2, and so on.

The second-level clustering results under Simulation Design 1 are summarized in Table 3.7. Overall, the final subclusters were consistent with the temporal shapes of the beta kernels rather than simply treating each kernel as an isolated class. Kernels with similar temporal locations and shapes tended to be grouped together. For example, kernels 1 and 12 both represent decreasing patterns starting from early pseudotime and were grouped into subcluster 1_1. Similarly, kernels 3 and 14 both capture early-peaking expression patterns and were grouped into subcluster 1_2.

For the middle temporal region, kernel 6 and kernel 17 were separated into two different subclusters, reflecting their distinct peak locations and widths. In the late temporal region, kernels 8 and 19 were also separated, while kernels 10 and 21 were grouped together because both represent late-peaking or increasing expression patterns near the end of pseudotime. These results suggest that the hierarchical sub-clustering procedure captured meaningful functional similarities among the fitted beta-kernel features and was able to recover both broad temporal trends and finer differences in peak timing.

Table 3.7: Second-level clustering results under Simulation Design 1. Rows represent the true kernel used in the simulation, and columns represent the final subclusters obtained after hierarchical sub-clustering.

True Kernel	**1_1**	**1_2**	**2_1**	**2_2**	**3_1**	**3_2**	**3_3**
1	52	0	0	0	0	0	0
12	48	0	0	0	0	0	0
3	0	52	0	0	0	0	0
14	0	48	0	0	0	0	0
6	0	0	52	0	0	0	0
17	0	0	0	48	0	0	0
8	0	0	0	0	52	0	0
19	0	0	0	0	0	48	0
10	0	0	0	0	0	0	52
21	0	0	0	0	0	0	48

Under Simulation Design 2, the second-level clustering results are shown in Table 3.8. The final subclusters again largely reflected the temporal shapes of the true beta kernels. Kernels 1 and 12, which represent decreasing expression patterns near the beginning of pseudotime, were grouped together in subcluster 1_1. Kernels 3 and 14, both corresponding to early-peaking patterns, were grouped together in subcluster 1_2.

For the middle and late temporal patterns, the clustering captured most of the expected structure but also showed some finer separation. Kernel 17 was mainly assigned to subcluster 2_1, whereas kernel 6 was primarily assigned to subcluster 2_2, with only one gene from kernel 6 assigned to 2_1. Kernel 8 was split between subclusters 3_1 and 3_2, while kernel 19 was assigned to 3_2. Finally, kernels 10 and 21 were grouped together in subcluster 3_3, consistent with their similar late-peaking or increasing temporal patterns. These results suggest that the sub-clustering

procedure remained robust under heterogeneous baseline expression levels, while also reflecting subtle differences among kernels with nearby temporal shapes.

Table 3.8: Second-level clustering results under Simulation Design 2. Rows represent the true kernel used in the simulation, and columns represent the final subclusters obtained after hierarchical sub-clustering.

True Kernel	**1_1**	**1_2**	**2_1**	**2_2**	**3_1**	**3_2**	**3_3**
1	52	0	0	0	0	0	0
12	48	0	0	0	0	0	0
3	0	52	0	0	0	0	0
14	0	48	0	0	0	0	0
6	0	0	1	51	0	0	0
17	0	0	48	0	0	0	0
8	0	0	0	0	38	14	0
19	0	0	0	0	0	48	0
10	0	0	0	0	0	0	52
21	0	0	0	0	0	0	48

Taken together, the functional clustering results from both simulation designs demonstrate that the fitted kernel-based features effectively capture the underlying temporal expression patterns at the gene-clustering level. In both designs, kernels with similar temporal shapes were consistently grouped together, while kernels with distinct temporal trajectories were well separated. The first-level clustering identified broad temporal structures, and the second-level clustering further resolved finer differences in peak location and curve shape within these broad groups.

Compared with Simulation Design 1, Simulation Design 2 showed slightly more variability in the second-level clustering, with a small number of genes assigned to neighboring subclusters. Nevertheless, the overall temporal organization was well preserved, and the resulting clusters remained highly consistent with the underlying

beta-kernel characteristics. These findings indicate that the proposed kernel-based features provide a robust representation of temporal gene expression dynamics and can recover biologically meaningful functional gene modules under both fixed and heterogeneous baseline expression settings.

3.5 Validation Using Known Time-Point Data

To further evaluate whether the proposed beta-kernel functions can capture biologically meaningful temporal expression patterns, we applied the kernel-based procedure to an independent malaria time-course dataset with known experimental collection times.[69] The Painter-pf dataset was generated from *Plasmodium falciparum* 3D7 parasites measured across the 48-hour intraerythrocytic developmental cycle (IDC), with samples collected hourly from 0 to 47 hours post-invasion (hpi). Unlike the single-cell RNA-seq datasets used in the main analysis, this dataset is not a raw count matrix. It is also not a CPM-normalized RNA-seq dataset. Instead, the expression profiles were obtained from DNA microarray experiments and further processed to produce normalized expression values.

The Painter-pf dataset contains three types of temporal expression profiles: total mRNA abundance, nascent transcription, and mRNA stabilization. These profiles describe complementary aspects of gene expression dynamics during the IDC. In the main validation analysis, we focused on the nascent transcription profile because it directly reflects active gene transcription over developmental time.

Because the true experimental collection time is available for each sample, this dataset provides a useful reference for validating whether the selected beta kernels can recover the timing of temporal gene expression patterns. For each gene, the observed peak time was obtained directly from the original transcription time-course by identifying the hpi at which the normalized expression value reached its maximum. This observed peak time was treated as the reference timing of the gene's temporal transcriptional activity.

We then applied the beta-kernel fitting and selection procedure to each gene using the 22 candidate beta kernels. The candidate kernels were divided into two sets. Set 1 contains sharper and more localized kernels, whereas Set 2 contains broader and smoother kernels. These two sets were designed to capture temporal expression patterns with different degrees of localization across the IDC. For each gene, the best-fitting kernel was selected based on the smallest AIC value. When multiple kernels produced the same best fit, kernels with a positive association with the expression profile were retained so that the selected kernel represented an activated temporal expression pattern rather than an inverse pattern. The peak location of the selected beta kernel was then rescaled from the unit interval to the 48-hour IDC scale and used as the estimated peak time for that gene.

Finally, we compared the estimated peak time from the selected beta kernel with the observed peak time obtained directly from the Painter-pf transcription profile. As shown in Figure 3.5, the estimated peak times were highly consistent with the observed peak times, with a Pearson correlation of 0.945. Since the estimated peak time is determined by the peak location of one of the 22 candidate kernels, the predicted peak times appear as discrete horizontal bands. Despite this discretization, most genes were distributed close to the diagonal reference line, indicating strong agreement between the kernel-estimated timing and the experimentally observed timing.

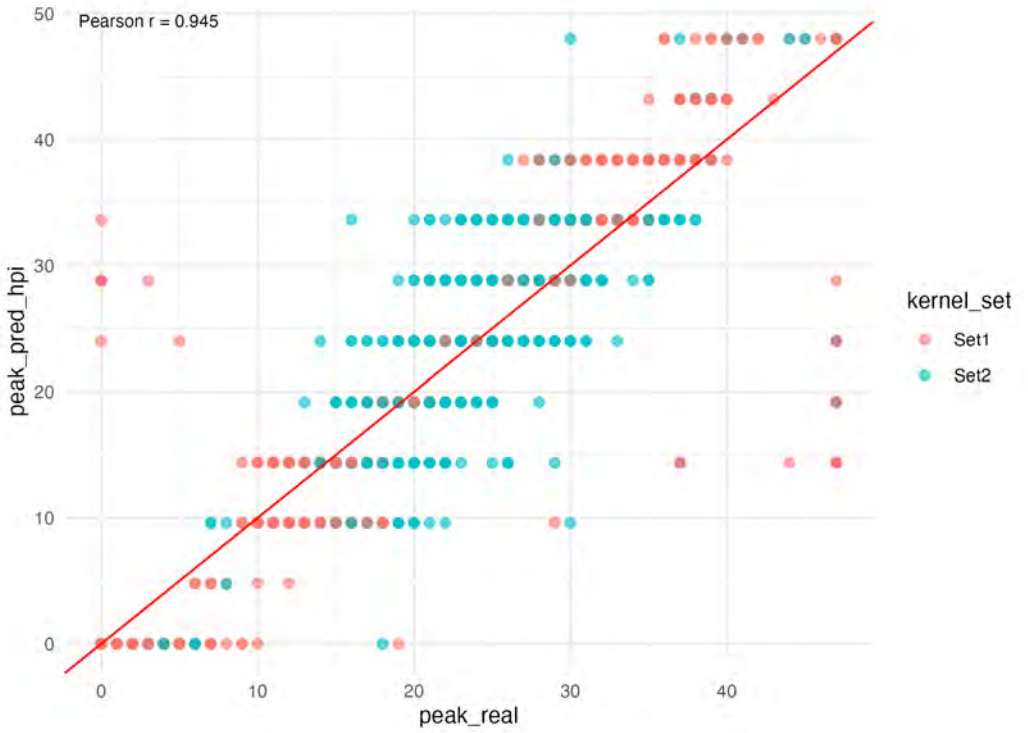


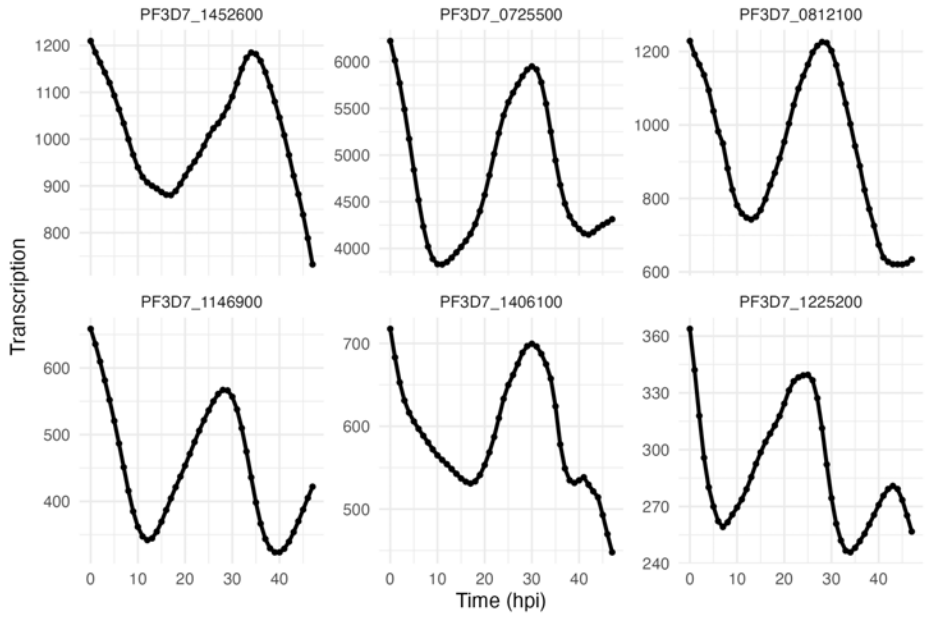
Figure 3.5: Validation of beta-kernel peak timing using the Painter-pf transcription time-course data. The x-axis shows the observed peak time, and the y-axis shows the estimated peak time from the selected beta kernel. Points are colored by kernel set, where Set 1 contains sharper kernels and Set 2 contains broader kernels. The red diagonal line indicates perfect agreement.

Although the overall agreement was strong, several off-diagonal genes were observed, particularly among genes with observed peak times at the boundaries of the IDC, such as 0 hpi and 47 hpi. To better understand these deviations, we examined representative off-diagonal genes with boundary peak times. As shown in Figure 3.6, many of these genes showed complex transcription profiles with multiple local peaks. For genes with an observed peak at 0 hpi, the expression was highest at the beginning of the time course but often showed another strong local peak later in the IDC. Similarly, for genes with an observed peak at 47 hpi, some profiles increased toward the end of the IDC but also contained earlier local activation patterns.

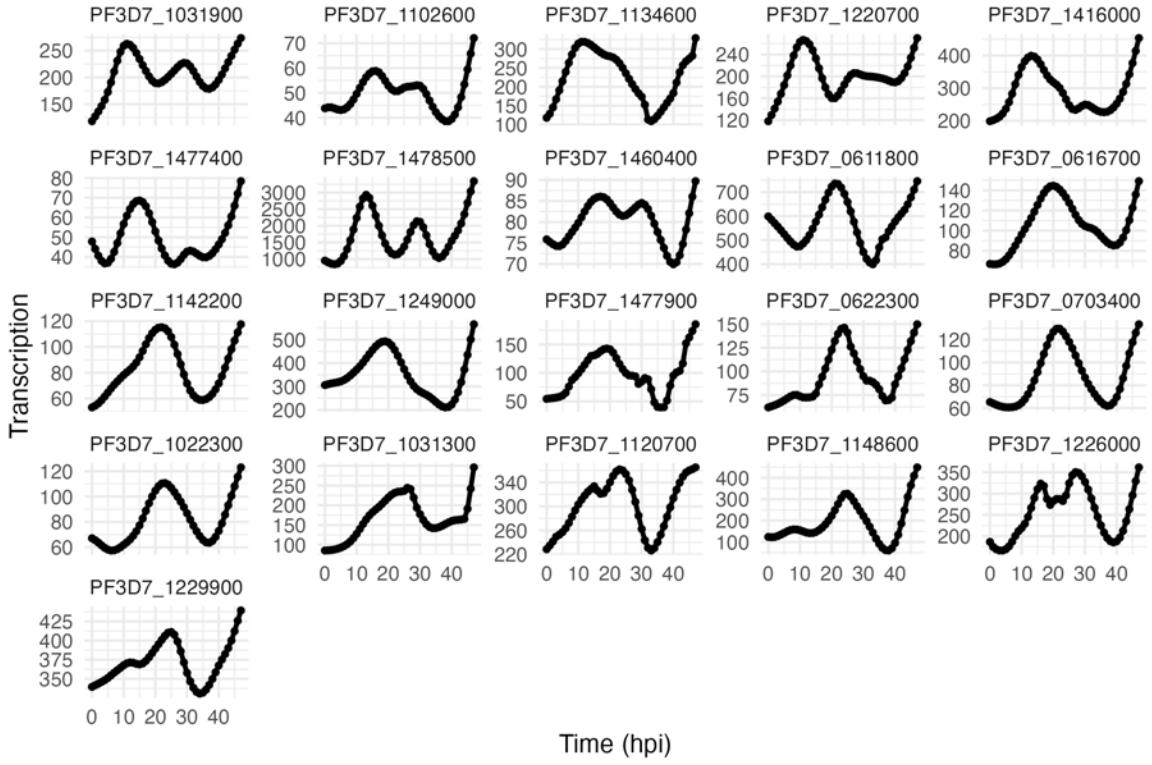
These discrepancies are likely due to two related factors. First, 0 hpi and 47 hpi are boundary points in the observed time scale, but biologically they are adjacent in the

cyclic IDC process. Therefore, a peak near the end of the cycle and a peak near the beginning of the cycle may appear far apart in a linear peak-time comparison. Second, the beta kernels are designed to capture a dominant unimodal temporal activation pattern. Genes with multiple peaks or boundary-driven maxima may not be fully represented by a single beta kernel. Thus, these off-diagonal genes do not necessarily indicate failure of the kernel library; rather, they highlight cases where the observed expression profile contains cyclic boundary effects or multiple temporal activation components.

Overall, these results suggest that the proposed beta-kernel library can effectively capture temporal transcriptional activation patterns across the malaria IDC, especially for genes with a dominant unimodal expression pattern. The Painter-pf validation therefore provides additional support for applying these kernels to malaria single-cell data, where the true biological time is unknown and must be approximated by pseudotime.



(A) Off-diagonal genes with real peak at 0 hpi.



(B) Off-diagonal genes with real peak at 47 hpi.

Figure 3.6: Examples of off-diagonal genes in the Painter-pf transcription validation. Some genes have boundary peaks at 0 or 47 hpi but also show additional local peaks during the IDC, which may lead the selected beta kernel to capture an internal activation pattern rather than the boundary maximum.

3.6 Real Single-Cell Data Application

After evaluating the performance of BetaDE in simulation studies, we applied the method to real single-cell RNA-seq data from *Plasmodium falciparum*. The goal of this real data analysis was to examine whether BetaDE could identify temporally dynamic genes and recover biologically meaningful expression patterns along the intraerythrocytic developmental cycle (IDC). Unlike the simulation studies, the true data-generating model and true temporal kernels are unknown in real data. Therefore, the evaluation focused on whether the selected kernels, estimated peak locations, and functional gene clusters were consistent with the known developmental progression of malaria parasites.

We first applied BetaDE to the Dogg *P. falciparum* single-cell RNA-seq dataset[58]. After preprocessing and quality control, the dataset contained 4,923 genes and 29,216 cells, including 7,499 ring-stage cells, 13,558 trophozoite-stage cells, and 8,159 schizont-stage cells. The three annotated stages represent the major developmental stages of the intraerythrocytic developmental cycle (IDC), but the transcriptional changes during this process are expected to occur continuously rather than as fully discrete states.

To examine the overall structure of the data, we first performed principal component analysis (PCA) on the processed single-cell expression matrix. As shown in Figure 3.7, cells formed a clear continuous trajectory in the first two principal components. Ring-stage cells were mainly located on one side of the trajectory, schizont-stage cells were located on the opposite side, and trophozoite-stage cells were broadly distributed between them. This pattern is consistent with the expected developmental progression from ring to trophozoite to schizont stages. At the same time, the overlap between neighboring stages indicates that the IDC progression is not fully captured by discrete stage labels. Therefore, a continuous pseudotime representation is more appropriate for modeling dynamic gene expression in this dataset.

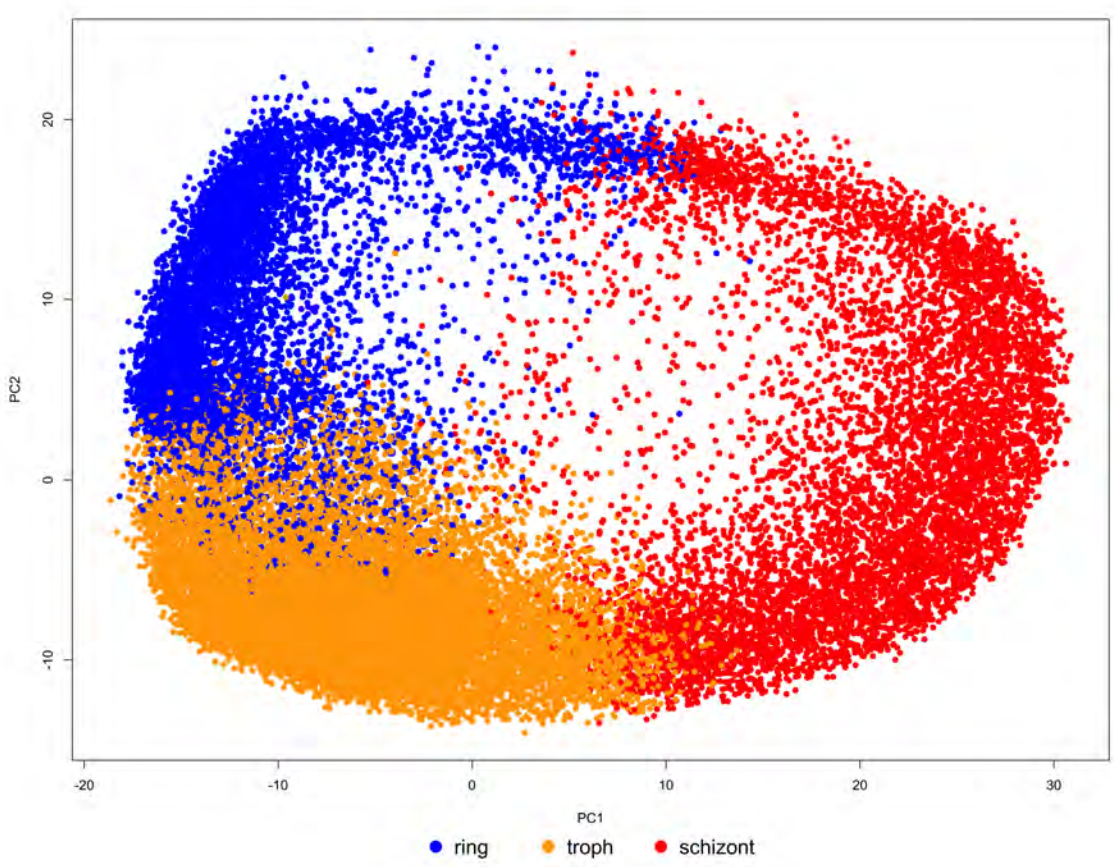


Figure 3.7: PCA plot of the Dogga *P. falciparum* single-cell RNA-seq dataset. Cells are colored by annotated IDC stage. The PCA result shows a continuous developmental structure from ring to trophozoite to schizont stages, supporting the use of pseudotime-based temporal modeling.

After confirming the continuous structure of the Dogga-pf dataset in PCA space, we next used Slingshot to infer a continuous developmental trajectory. The annotated stage information was used to guide the lineage direction from ring to trophozoite to schizont stages. The resulting pseudotime values were rescaled to the interval $(0, 1)$, with smaller values corresponding to early IDC stages and larger values corresponding to late IDC stages.

Using the inferred pseudotime from Slingshot, we applied BetaDE to model temporal gene expression patterns in the Dogga-pf dataset. For each gene, BetaDE fitted Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial

models across the candidate beta kernels. The best model-kernel combination was selected based on AIC, and temporal differential expression was assessed by testing whether the selected kernel effect was significantly different from zero.

The selected kernels provided an interpretable summary of the activation timing of each gene along the IDC. Genes assigned to early-peaking kernels showed stronger expression near the beginning of pseudotime, corresponding mainly to ring-stage cells. Genes assigned to intermediate kernels tended to peak in the trophozoite region, while genes assigned to late-peaking kernels were more active near the schizont stage. Therefore, the selected beta kernels captured stage-specific and stage-transition expression patterns in a continuous pseudotime framework.

The model selection results also reflected the heterogeneity of single-cell count data. Some genes were best fitted by Poisson or negative binomial models, while others were assigned to zero-inflated models. This suggests that different genes exhibit different levels of overdispersion and sparsity, and that the flexible model-selection strategy in BetaDE is useful for adapting to gene-specific expression characteristics.

To further investigate the temporal structure among the significant genes, we performed functional clustering based on the fitted BetaDE features. For each gene, the feature vector was constructed using the fitted beta-kernel coefficients together with selected-kernel summary features, including the selected kernel location and variance. These features summarize both the strength and timing of the estimated temporal expression pattern.

Hierarchical clustering was first applied to the standardized BetaDE feature matrix to identify broad temporal gene modules. As shown in Figure 3.8, the significant temporally dynamic genes were separated into three major first-level clusters. These broad clusters indicate that the BetaDE-derived features captured large-scale differences in temporal expression patterns among genes.

However, the dendrogram also shows that each major cluster still contains multi-

ple internal branches. This suggests that genes within the same broad cluster may share similar overall temporal behavior but still differ in finer aspects, such as activation timing, peak width, or expression shape along pseudotime. Therefore, we further performed sub-clustering within each first-level cluster. This two-level clustering strategy allowed us to capture both broad temporal modules and more detailed temporal sub-patterns among dynamic genes.

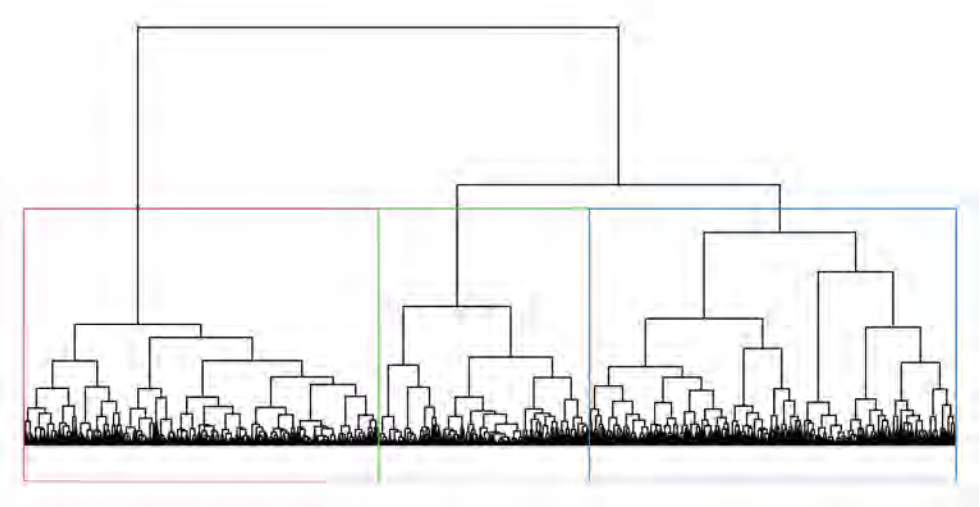


Figure 3.8: First-level hierarchical clustering of significant temporally dynamic genes in the Dogga-pf dataset. Genes were clustered using standardized BetaDE-derived features. The dendrogram was cut into three broad clusters, shown by the colored rectangles.

To assess whether the gene programs recovered by BetaDE correspond to biologically meaningful units, we performed GO biological-process enrichment analysis for each functional subcluster derived from the Dogga-pf dataset (Figure 3.9). BetaDE partitioned genes into three top-level temporal clusters, and each top-level cluster was further divided into functional subclusters, denoted by the format $\text{cluster}_{\text{subcluster}}$. For example, (2_1) represents the first subcluster within top-level cluster 2. In Figure 3.9, each point represents a significantly enriched GO biological-process term within a subcluster after Benjamini–Hochberg correction with adjusted ($p < 0.05$). Point size indicates the percentage of genes in the subcluster annotated to the corresponding

GO term, while point color represents the enrichment significance, with red indicating stronger significance and blue indicating weaker significance. The number in parentheses following each GO term indicates the total number of genes annotated to that term.

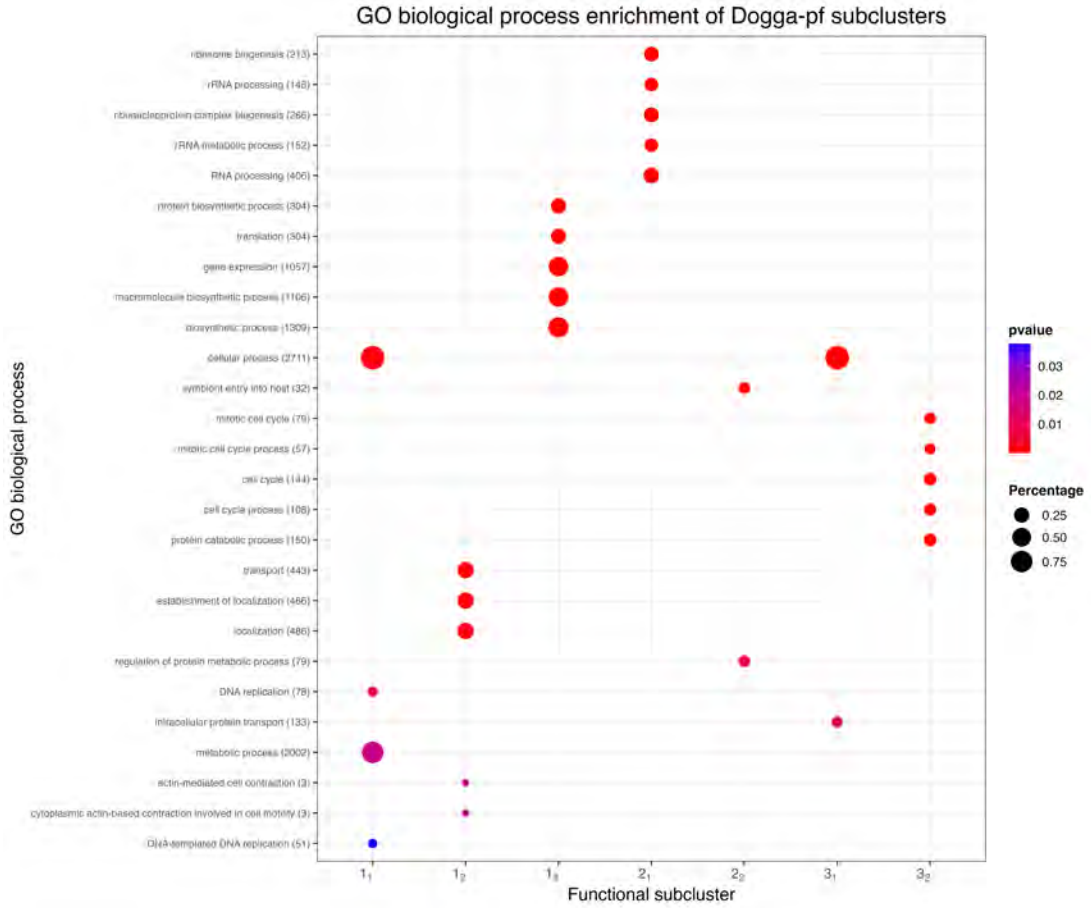


Figure 3.9: GO biological process enrichment analysis of BetaDE-derived functional subclusters in the Dogga-pf dataset. Each point represents an enriched GO biological process term within a functional subcluster. The point size indicates the percentage of genes in the subcluster associated with the GO term, and the color indicates the enrichment p-value.

Each subcluster was enriched for a distinct and internally coherent set of processes, with little overlap between subclusters, indicating that BetaDE partitions genes into functionally specific modules rather than arbitrary groupings.

The enriched programs recapitulated the principal functional transitions of the

intraerythrocytic developmental cycle. Subcluster 2_1 was strongly enriched for ribosome biogenesis, rRNA processing, ribonucleoprotein complex biogenesis, and RNA processing, while subcluster 1_3 was enriched for translation, protein biosynthesis, and gene expression. Together, these subclusters define a protein-synthesis program consistent with the metabolically active trophozoite stage, during which the parasite substantially upregulates ribosome assembly and translational activity. Subcluster 1_2 was enriched for transport and localization processes, reflecting the intensified macromolecular trafficking that accompanies this growth phase.

A separate set of subclusters captured the processes of parasite replication and host invasion. Subcluster 3_2 was enriched for mitotic cell cycle, cell cycle, and protein catabolic processes, consistent with schizogonyâthe phase of DNA replication and nuclear division that produces daughter merozoites. Subcluster 2_2 was enriched for "symbiont entry into host" and the regulation of protein metabolism, corresponding to the invasion program expressed as mature schizonts give rise to invasive merozoites. Subcluster 1_1 was dominated by broad metabolic and DNA-replication processes, including DNA replication and DNA-templated DNA replication; the enriched terms in this subcluster were the most general (e.g., cellular process, metabolic process) and showed the highest gene percentages but more modest enrichment significance, consistent with a program of housekeeping and basal metabolic activity rather than a stage-restricted function.

Across all subclusters, the enriched terms were biologically consistent with the developmental program of the asexual blood stages, progressing from basal metabolism and macromolecular synthesis through translation and growth to cell-cycle division and host invasion. The recovery of these stage-characteristic, functionally segregated programsâwithout any prior stage information supplied to the methodâindicates that BetaDE identifies temporal gene programs that reflect genuine biological structure rather than computational artifacts.

3.7 Discussion

In this project, we developed BetaDE, a beta-kernel-based framework for modeling pseudotime-dependent gene expression patterns in single-cell malaria transcriptomic data. By representing temporal expression dynamics using a set of beta kernels, BetaDE provides an interpretable way to capture early, middle, late, and transient activation patterns during the intraerythrocytic developmental cycle. In addition, by considering multiple count-based distributions, including Poisson, negative binomial, ZIP, and ZINB models, the framework accounts for both count variability and potential excess zeros in single-cell RNA-seq data.

The simulation results showed that the fitted beta-kernel features were able to recover the underlying temporal gene modules. Although the first-level clustering mainly captured broad temporal structures, the second-level clustering further separated genes into finer subclusters that were largely consistent with the true kernel groups. These results suggest that the beta-kernel representation provides useful features for functional clustering of temporally dynamic genes.

We then applied BetaDE to the Dogga-pf single-cell dataset to evaluate its performance on real malaria transcriptomic data. The resulting temporal subclusters showed several biologically interpretable functional patterns. GO enrichment analysis identified ribosome biogenesis and rRNA processing in subcluster 2_1, translation and biosynthetic activity in subcluster 1_3, cell-cycle-related processes in subcluster 3_2, and transport/localization-related processes in subclusters 1_2 and 3_1. These results suggest that the BetaDE-derived temporal modules capture biologically relevant expression programs during parasite development.

However, some enriched GO terms were broad or shared across multiple subclusters, such as gene expression, biosynthetic process, metabolic process, and cellular process. Therefore, the GO enrichment results were interpreted as supportive evidence rather than definitive cluster-specific functional annotations. In addition, the performance

of BetaDE depends on the quality of pseudotime estimation and on whether the predefined beta-kernel dictionary adequately represents the true temporal expression patterns.

Overall, this project demonstrates that beta-kernel-based temporal modeling is a useful and interpretable approach for identifying dynamic gene expression modules in single-cell malaria data. The simulation and Dogga-pf real data results together suggest that BetaDE can recover meaningful temporal structures and provide biologically relevant functional insights.

CHAPTER 4: CONCLUSIONS AND FUTURE WORK

This dissertation developed computational methods for analyzing transcriptomic data from both bulk RNA-seq and single-cell RNA-seq perspectives. The two projects addressed complementary problems in transcriptomic studies of malaria parasite development. The first project focused on estimating underlying parasite stage compositions from bulk RNA-seq mixtures, where the observed expression profile represents a mixture of multiple developmental stages. The second project focused on modeling dynamic gene expression patterns along single-cell pseudotime and identifying temporal gene modules. Together, these studies demonstrate the value of integrating single-cell information, matrix factorization, temporal modeling, and functional clustering to improve the interpretation of parasite developmental dynamics. Although the two projects were designed for different data types and scientific questions, both aimed to extract biologically meaningful structure from complex and noisy transcriptomic data.

In Chapter 2, we proposed GSNMF+, a geometry-guided nonnegative matrix factorization framework with data augmentation for robust deconvolution of bulk RNA-seq data. By incorporating single-cell-derived geometric information through solvability and manifold regularization terms, GSNMF+ improves the biological interpretability and stability of bulk deconvolution. Simulation studies showed that GSNMF+ can recover underlying stage proportions under multiple mixture scenarios, and real data applications demonstrated its potential for estimating parasite developmental composition when true proportions are unknown. These results suggest that single-cell-guided geometric constraints and data augmentation can provide useful information for improving bulk RNA-seq deconvolution in malaria transcriptomic studies.

In Chapter 3, we developed BetaDE, a beta-kernel-based framework for modeling pseudotime-dependent gene expression dynamics in single-cell malaria data. BetaDE represents temporal expression patterns using interpretable beta kernels and accounts

for count variability and excess zeros through Poisson, negative binomial, ZIP, and ZINB models. The simulation studies showed that the fitted beta-kernel features can recover underlying temporal gene modules and separate genes with different activation patterns along pseudotime. In the Dogga-pf real data application, BetaDE identified functional subclusters associated with biologically meaningful processes, including ribosome biogenesis, rRNA processing, translation, biosynthetic activity, cell-cycle-related processes, and transport/localization. These results suggest that beta-kernel-based temporal modeling provides an interpretable way to identify dynamic gene expression programs during parasite development.

For future development, the current framework can be extended beyond fitting each gene using a single selected kernel. Instead, multiple beta kernels could be combined simultaneously to model more complex gene expression patterns, especially genes with broad, multi-phase, or multiple-peak temporal dynamics. This multi-kernel representation may provide a more flexible description of gene expression trajectories while still preserving interpretability. Future work could also explore alternative clustering strategies based on the fitted temporal models. For example, fitted expression curves or model-based temporal features could be clustered using functional principal component analysis (FPCA), fuzzy c-means (FCM), or other functional clustering approaches. Comparing these alternative clustering results with the two-level clustering results used in this dissertation would help assess the robustness of the identified temporal gene modules and determine whether similar biological patterns are recovered across different clustering frameworks.

REFERENCES

[1] D. Chen, S. Li, and X. Wang, "Geometric structure guided model and algorithms for complete deconvolution of gene expression data," *Foundations of Data Science*, vol. 4, no. 3, pp. 441–476, 2022.

[2] World Health Organization, "World malaria report 2025," tech. rep., World Health Organization, Geneva, 2025.

[3] S. Sato, "Plasmodium—a brief introduction to the parasites causing human malaria and their basic biology," *Journal of Physiological Anthropology*, vol. 40, no. 1, p. 1, 2021.

[4] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[5] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by rna-seq," *Nature Methods*, vol. 5, pp. 621–628, 2008.

[6] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[7] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome Biology*, vol. 15, p. 550, 2014.

[8] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani, "mrna-seq whole-transcriptome analysis of a single cell," *Nature Methods*, vol. 6, pp. 377–382, 2009.

[9] A.-E. Saliba, A. J. Westermann, S. A. Gorski, and J. Vogel, "Single-cell rna-seq: advances and future challenges," *Nucleic Acids Research*, vol. 42, no. 14, pp. 8845–8860, 2014.

[10] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll, "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets," *Cell*, vol. 161, no. 5, pp. 1202–1214, 2015.

[11] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H.

Bielas, "Massively parallel digital transcriptional profiling of single cells," *Nature Communications*, vol. 8, p. 14049, 2017.

[12] A. J. Reid, A. M. Talman, H. M. Bennett, A. R. Gomes, M. J. Sanders, C. J. R. Illingworth, O. Billker, M. Berriman, and M. K. N. Lawniczak, "Single-cell rna-seq reveals hidden transcriptional variation in malaria parasites," *eLife*, vol. 7, p. e33105, 2018.

[13] V. M. Howick, A. J. C. Russell, T. Andrews, H. Heaton, A. J. Reid, K. Natarajan, H. Butungi, T. Metcalf, L. H. Verzier, J. C. Rayner, M. Berriman, J. K. Herren, O. Billker, M. Hemberg, A. M. Talman, and M. K. N. Lawniczak, "The malaria cell atlas: single parasite transcriptomes across the complete plasmodium life cycle," *Science*, vol. 365, no. 6455, p. eaaw2619, 2019.

[14] M. D. Luecken and F. J. Theis, "Current best practices in single-cell rna-seq analysis: a tutorial," *Molecular Systems Biology*, vol. 15, no. 6, p. e8746, 2019.

[15] V. Y. Kiselev, T. S. Andrews, and M. Hemberg, "Challenges in unsupervised clustering of single-cell rna-seq data," *Nature Reviews Genetics*, vol. 20, no. 5, pp. 273–282, 2019.

[16] P. Qiu, "Embracing the dropouts in single-cell rna-seq analysis," *Nature Communications*, vol. 11, p. 1169, 2020.

[17] A. Haque, J. Engel, S. A. Teichmann, and T. Lönnberg, "A practical guide to single-cell rna-sequencing for biomedical research and clinical applications," *Genome Medicine*, vol. 9, no. 1, p. 75, 2017.

[18] F. Avila Cobos, J. Alquicira-Hernandez, J. E. Powell, P. Mestdagh, and K. De Preter, "Benchmarking of cell type deconvolution pipelines for transcriptomics data," *Nature Communications*, vol. 11, p. 5650, 2020.

[19] A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, and A. A. Alizadeh, "Robust enumeration of cell subsets from tissue expression profiles," *Nature Methods*, vol. 12, no. 5, pp. 453–457, 2015.

[20] X. Wang, J.-A. Park, K. Susztak, N. R. Zhang, and M. Li, "Bulk tissue cell type deconvolution with multi-subject single-cell expression reference," *Nature Communications*, vol. 10, p. 380, 2019.

[21] R. Joice, V. Narasimhan, J. Montgomery, A. B. Sidhu, K. Oh, E. Meyer, W. Pierre-Louis, K. Seydel, D. Milner, K. Williamson, R. Wiegand, D. Ndiaye, J. Daily, D. Wirth, T. Taylor, C. Huttenhower, and M. Marti, "Inferring developmental stage composition from gene expression in human malaria," *PLOS Computational Biology*, vol. 9, no. 12, p. e1003392, 2013.

[22] K. Tebben, A. Dia, and D. Serre, "Determination of the stage composition of *Plasmodium* infections from bulk gene expression data," *mSystems*, vol. 7, no. 4, pp. e00258–22, 2022.

[23] A. M. Newman, C. B. Steen, C. L. Liu, A. J. Gentles, A. A. Chaudhuri, F. Scherer, M. S. Khodadoust, M. S. Esfahani, B. A. Luca, D. Steiner, M. Diehn, and A. A. Alizadeh, "Determining cell type abundance and expression from bulk tissues with digital cytometry," *Nature Biotechnology*, vol. 37, pp. 773–782, 2019.

[24] K. Menden, M. Marouf, S. Oller, A. Dalmia, D. S. Magruder, K. Kloiber, P. Heutink, and S. Bonn, "Deep learning-based cell composition analysis from tissue expression profiles," *Science Advances*, vol. 6, no. 30, p. eaba2619, 2020.

[25] T. Chu, Z. Wang, D. Pe'er, and C. G. Danko, "Cell type and gene expression deconvolution with bayesprism enables bayesian integrative analysis across bulk and single-cell rna sequencing in oncology," *Nature Cancer*, vol. 3, no. 4, pp. 505–517, 2022.

[26] K. Kang, Q. Meng, I. Shats, D. M. Umbach, M. Li, Y. Li, X. Li, and L. Li, "Cdseq: A novel complete deconvolution method for dissecting heterogeneous samples using gene expression data," *PLOS Computational Biology*, vol. 15, no. 12, p. e1007510, 2019.

[27] B. F. Miller, F. Huang, L. Atta, A. Sahoo, and J. Fan, "Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data," *Nature Communications*, vol. 13, p. 2339, 2022.

[28] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications," *arXiv preprint arXiv:1803.01257*, 2018.

[29] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1336–1353, 2012.

[30] N. Gillis, *Nonnegative Matrix Factorization: Complexity, Algorithms and Applications*. PhD thesis, Université catholique de Louvain, 2011.

[31] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?," in *Advances in Neural Information Processing Systems*, vol. 16, 2003.

[32] K. Huang, N. D. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 211–224, 2014.

[33] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen, "Theorems on positive data: On the uniqueness of nmf," *Computational Intelligence and Neuroscience*, vol. 2008, p. 764206, 2008.

[34] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells," *Nature Biotechnology*, vol. 32, no. 4, pp. 381–386, 2014.

[35] K. Van den Berge, H. Roux de Bézieux, K. Street, W. Saelens, R. Cannoodt, Y. Saeys, S. Dudoit, and L. Clement, "Trajectory-based differential expression analysis for single-cell sequencing data," *Nature Communications*, vol. 11, no. 1, p. 1201, 2020.

[36] Z. Bozdech, M. Llinás, B. L. Pulliam, E. D. Wong, J. Zhu, and J. L. DeRisi, "The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum," *PLoS Biology*, vol. 1, no. 1, p. e5, 2003.

[37] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J.-P. Vert, "A general and flexible method for signal extraction from single-cell rna-seq data," *Nature Communications*, vol. 9, p. 284, 2018.

[38] A. Sarkar and M. Stephens, "Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis," *Nature Genetics*, vol. 53, pp. 770–777, 2021.

[39] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys, "A comparison of single-cell trajectory inference methods," *Nature Biotechnology*, vol. 37, no. 5, pp. 547–554, 2019.

[40] L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, and F. J. Theis, "Diffusion pseudotime robustly reconstructs lineage branching," *Nature Methods*, vol. 13, no. 10, pp. 845–848, 2016.

[41] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit, "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics," *BMC Genomics*, vol. 19, no. 1, p. 477, 2018.

[42] F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, and F. J. Theis, "Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells," *Genome Biology*, vol. 20, no. 1, p. 59, 2019.

[43] D. Song and J. J. Li, "Pseudotimede: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell rna sequencing data," *Genome Biology*, vol. 22, no. 1, p. 124, 2021.

[44] P.-F. Kuan and X. Ren, "Negative binomial additive model for rna-seq data analysis," *BMC Bioinformatics*, vol. 21, p. 171, 2020.

[45] D. S. Fischer, F. J. Theis, and N. Yosef, "Impulse model-based differential expression analysis of time course sequencing data," *Nucleic Acids Research*, vol. 46, no. 20, p. e119, 2018.

[46] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*. New York: Springer, 2 ed., 2005.

[47] J.-L. Wang, J.-M. Chiou, and H.-G. Müller, "Functional data analysis," *Annual Review of Statistics and Its Application*, vol. 3, pp. 257–295, 2016.

[48] J. J. Song, H.-J. Lee, J. S. Morris, and S. Kang, "Clustering of time-course gene expression data using functional data analysis," *Computational Biology and Chemistry*, vol. 31, no. 4, pp. 265–274, 2007.

[49] L. Kumar and M. E. Futschik, "Mfuzz: a software package for soft clustering of microarray data," *Bioinformation*, vol. 2, no. 1, pp. 5–7, 2007.

[50] J. Ernst and Z. Bar-Joseph, "Stem: a tool for the analysis of short time series gene expression data," *BMC Bioinformatics*, vol. 7, p. 191, 2006.

[51] M. E. Futschik and B. Carlisle, "Noise-robust soft clustering of gene expression time-course data," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 4, pp. 965–988, 2005.

[52] Y.-H. Luan and H. Li, "Clustering of time-course gene expression data using a mixed-effects model with b-splines," *Bioinformatics*, vol. 19, no. 4, pp. 474–482, 2003.

[53] L. A. Newberg, X. Chen, C. D. Kodira, and M. I. Zavodszky, "Computational de novo discovery of distinguishing genes for biological processes and cell types in complex tissues," *PLOS ONE*, vol. 13, no. 2, p. e0193067, 2018.

[54] K. Zaitsev, M. Bambouskova, A. Swain, and M. N. Artyomov, "Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures," *Nature Communications*, vol. 10, no. 1, p. 2209, 2019.

[55] Z. Li and H. Wu, "TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis," *Genome Biology*, vol. 20, no. 1, p. 190, 2019.

[56] L. Price, T. Planche, C. Rayner, and S. Krishna, "Acute respiratory distress syndrome in Plasmodium vivax malaria: case report and review of the literature," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 101, no. 7, pp. 655–659, 2007.

[57] N. M. Anstey, B. Russell, T. W. Yeo, and R. N. Price, "The pathophysiology of Plasmodium vivax malaria," *Trends in Parasitology*, vol. 25, no. 5, pp. 220–227, 2009.

[58] S. K. Dogga, J. C. Rop, J. Cudini, E. Farr, A. Dara, D. Ouologuem, A. A. Djimdé, A. M. Talman, and M. K. N. Lawniczak, "A single cell atlas of sexual development in Plasmodium falciparum," *Science*, vol. 384, no. 6695, p. eadj4088, 2024.

[59] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[60] D. D. Lee and H. S. Seung, "Algorithms for Non-negative Matrix Factorization," in *Advances in Neural Information Processing Systems 13*, pp. 556–562, Neural Information Processing Systems Foundation, 2001.

[61] A. Kim, J. Popovici, D. Menard, and D. Serre, "Plasmodium vivax transcriptomes reveal stage-specific chloroquine response and differential regulation of male and female gametocytes," *Nature Communications*, vol. 10, no. 1, p. 371, 2019.

[62] D. Kepple, C. T. Ford, J. Williams, B. Abagero, S. Li, J. Popovici, D. Yewhalaw, and E. Lo, "Comparative transcriptomics reveal differential gene expression among Plasmodium vivax geographical isolates and implications on erythrocyte invasion mechanisms," *PLOS Neglected Tropical Diseases*, vol. 18, no. 1, p. e0011926, 2024.

[63] Y. Xin, J. Kim, H. Okamoto, M. Ni, Y. Wei, C. Adler, A. J. Murphy, G. D. Yancopoulos, C. Lin, and J. Gromada, "RNA sequencing of single human islet cells reveals type 2 diabetes genes," *Cell Metabolism*, vol. 24, no. 4, pp. 608–615, 2016. GEO accession: GSE81608.

[64] Å. Segerstolpe, A. Palasantza, P. Eliasson, E.-M. Andersson, A.-C. Andr'easson, X. Sun, S. Picelli, A. Sabirsh, M. Clausen, M. K. Bjursell, D. M. Smith, M. Kasper, C. Ammala, and R. Sandberg, "Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes," *Cell Metabolism*, vol. 24, no. 4, pp. 593–607, 2016. ArrayExpress accession: E-MTAB-5061.

[65] G. Atla, S. Bon'as-Guarch, M. Cuenca-Ardura, A. Beucher, D. J. M. Crouch, J. Garcia-Hurtado, I. Moran, M. Irimia, R. B. Prasad, A. L. Gloyn, *et al.*, "Genetic regulation of RNA splicing in human pancreatic islets," *Genome Biology*, vol. 23, p. 196, 2022.

[66] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.

[67] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey, "The sva package for removing batch effects and other unwanted variation in high-throughput experiments," *Bioinformatics*, vol. 28, no. 6, pp. 882–883, 2012.

[68] K. P. Burnham and D. R. Anderson, "Multimodel inference: Understanding AIC and BIC in model selection," *Sociological Methods & Research*, vol. 33, no. 2, pp. 261–304, 2004.

[69] H. J. Painter, N. C. Chung, A. Sebastian, I. Albert, J. D. Storey, and M. Llinás, "Genome-wide real-time in vivo transcriptional dynamics during Plasmodium falciparum blood-stage development," *Nature Communications*, vol. 9, p. 2656, 2018.

APPENDIX A: Derivation of Multiplicative Update Rules for GSNMF+

This appendix provides the derivation of the multiplicative update rules used in GSNMF+. Throughout this appendix, i and ℓ index genes, j and r index latent components, cell types, or developmental stages, and s indexes bulk or pseudo-bulk samples. Let

$$\mathbf{G} \in \mathbb{R}_+^{N\times n}, \qquad \mathbf{C} \in \mathbb{R}_+^{N\times k}, \qquad \mathbf{P} \in \mathbb{R}_+^{k\times n},$$

where N is the number of genes, n is the number of bulk or pseudo-bulk samples, and k is the number of latent components.

The full GSNMF+ objective function is

$$\mathcal{F}(\mathbf{C},\mathbf{P}) = \frac{1}{2}\|\mathbf{G}-\mathbf{CP}\|_F^2 + F_1(\mathbf{C}) + F_2(\mathbf{C}) + \mathbf{1}_T(\mathbf{P}), \tag{A.1}$$

where $F_1(\mathbf{C})$ is the solvability penalty, $F_2(\mathbf{C})$ is the manifold regularization penalty, and $\mathbf{1}_T(\mathbf{P})$ is the indicator function for the simplex constraint

$$T = \left\{ \mathbf{P} \in \mathbb{R}_+^{k\times n} : \sum_{r=1}^{k} P_{rs} = 1,\ s = 1,\dots,n \right\}.$$

A.1 Solvability Penalty

Let S_r denote the set of marker genes assigned to component r, for $r = 1,\dots,k$. Define the function $g(i)$ as

$$g(i) = r \quad \text{if } i \in S_r,$$

and set $g(i) = 0$ for non-marker genes. We also define the indicator function

$$\chi_i = \begin{cases} 1, & g(i) \neq 0, \\ 0, & g(i) = 0. \end{cases}$$

The solvability penalty encourages marker-gene rows of $\mathbf{C}$ to align with the corresponding coordinate direction. It is defined as

$$\begin{aligned} F_1(\mathbf{C}) &= \frac{\lambda_1}{2} \sum_{r=1}^{k} \sum_{i \in S_r} d_{\text{Eisen}} \left(\mathbf{C}_{(i)}, \mathbf{e}_r^\top \right)^2 \\ &= \frac{\lambda_1}{2} \sum_{r=1}^{k} \sum_{i \in S_r} \left(1 - \frac{\langle \mathbf{C}_{(i)}, \mathbf{e}_r^\top \rangle}{|\mathbf{C}_{(i)}||\mathbf{e}_r^\top|} \right)^2 \\ &= \frac{\lambda_1}{2} \sum_{r=1}^{k} \sum_{i \in S_r} \left(1 - \frac{C_{ig(i)}}{|\mathbf{C}_{(i)}|} \right)^2 . \end{aligned} \tag{A.2}$$

For each gene i, define the row-wise ℓ_2-norm of $\mathbf{C}$ as

$$|C_{(i)}| = \sqrt{\sum_{j=1}^{k} c_{ij}^2} = \left(\sum_{j=1}^{k} c_{ij}^2 \right)^{\frac{1}{2}}$$

Then

$$|C_{(i)}|' = \frac{1}{2} \left(\sum_{j=1}^{k} c_{ij}^2 \right)^{-\frac{1}{2}} \cdot 2c_{ij} = \frac{c_{ij}}{|C_{(i)}|} \tag{A.3}$$

Therefore, the gradient of $\mathcal{F}_1(C)$:

$$\frac{\partial \mathcal{F}_1}{\partial c_{ij}} = \lambda_1 \sum_{i \in S_r} \left(1 - \frac{C_{ig(i)}}{|C_{(i)}|} \right) \cdot (-1) \cdot \frac{\delta_{g(i),j} \cdot |C_{(i)}| - C_{ig(i)} \cdot \frac{c_{ij}}{|C_{(i)}|}}{|C_{(i)}|^2} \tag{A.4}$$

$$= \lambda_1 \sum_{i \in S_r} \left(1 - \frac{C_{ig(i)}}{|C_{(i)}|} \right) \cdot (-1) \cdot \left(\frac{\delta_{g(i),j}}{|C_{(i)}|} - \frac{C_{ig(i)} c_{ij}}{|C_{(i)}|^3} \right) \tag{A.5}$$

$$= \lambda_1 \sum_{i \in S_r} \left(1 - \frac{C_{ig(i)}}{|C_{(i)}|} \right) \cdot \frac{1}{|C_{(i)}|} \cdot \left(\frac{C_{ig(i)} c_{ij}}{|C_{(i)}|^2} - \delta_{g(i),j} \right) \tag{A.6}$$

$$= \lambda_1 \cdot \chi_{g(i)} \cdot \left(1 - \frac{C_{ig(i)}}{|C_{(i)}|} \right) \cdot \frac{1}{|C_{(i)}|} \cdot \left(\frac{C_{ig(i)} c_{ij}}{|C_{(i)}|^2} - \delta_{g(i),j} \right) \tag{A.7}$$

where $g : \mathcal{S}_r \to r$ is a surjective function, i.e., $g(i) = r$ if there is an r such that $i \in \mathcal{S}_r$ or $g(i) = 0$ otherwise. This function is known from the spectral clustering and $g(i) = r$ means gene i is the marker gene of the r-th cell type. To take into account all possible non-marker genes, we use the characteristic function $\chi_{g(i)} = 1$

if $g(i) \neq 0$ and $\chi_{g(i)} = 0$ otherwise. It is convenient to write Eq. A.7 in matrix form for future computation. To do so, we define: $\mathbf{W}_1 = \text{diag}\left\{\left(1 - \frac{c_{ig(i)}}{|\mathbf{C}_{(i)}|}\right)\frac{\chi_{g(i)}}{|\mathbf{C}_{(i)}|}\right\}$; $\mathbf{W}_2 = \text{diag}\left\{\frac{c_{ig(i)}}{|\mathbf{C}_{(i)}|^2}\right\}$, and $\mathbf{C}_g = \{\delta_{g(i),j}\}$ Here we use the notation $\text{diag}\{x_i\}$ to mean a diagonal matrix with all entries x_i on its diagonal, and $\text{diag}(\mathbf{x})$ means a vector $\mathbf{x}$ forms a diagonal matrix with the corresponding size. Then the matrix form of Eq (A.7) is:

$$\nabla \mathcal{F}_1 = \lambda_1 \mathbf{W}_1 \left(\mathbf{W}_2 \mathbf{C} - \mathbf{C}_g\right). \tag{A.8}$$

A.2 Manifold Regularization Penalty

Let $\omega_{i\ell}$ denote the similarity weight between genes i and ℓ. The manifold regularization penalty encourages genes that are close in the observed expression space to remain close in the latent component space. It is defined as

$$\begin{aligned} F_2(\mathbf{C}) &= \frac{\lambda_2}{2} \sum_{\ell=1}^{N} \sum_{i=1}^{N} \omega_{i\ell} d_{\text{eisen}}\left(\mathbf{C}_{(i)}, \mathbf{C}_{(\ell)}\right)^2 \\ &= \frac{\lambda_2}{2} \sum_{\ell=1}^{N} \sum_{i=1}^{N} \omega_{i\ell} \left(1 - \frac{\langle \mathbf{C}_{(i)}, \mathbf{C}_{(\ell)} \rangle}{|\mathbf{C}_{(i)}||\mathbf{C}_{(\ell)}|}\right)^2. \end{aligned} \tag{A.9}$$

For fixed i and ℓ, the gradient of numerator in Eq. A.9 is

$$\frac{\partial \langle C_{(i)}, C_{(l)} \rangle}{\partial c_{ij}} = \frac{\partial}{\partial c_{ij}} \left(\sum_{j=1}^{k} c_{ij} c_{lj}\right) \tag{A.10}$$

$$= c_{lj} \tag{A.11}$$

the gradient of denominator in Eq. A.9 is

$$\frac{\partial}{\partial c_{ij}}\left(|C_{(i)}||C_{(l)}|\right) = \frac{\partial}{\partial c_{ij}}\left(\sqrt{\sum_{j=1} c_{ij}^2}\cdot\sqrt{\sum_{j=1} c_{lj}^2}\right) \tag{A.12}$$

$$= \frac{\partial}{\partial c_{ij}}\left(\left(\sum_{j=1} c_{ij}^2\right)^{\frac{1}{2}}|C_{(l)}|\right) \tag{A.13}$$

$$= \frac{1}{2}\left(\sum_{j=1}^{k} c_{ij}^2\right)^{-\frac{1}{2}}\cdot 2c_{ij}\cdot|C_{(l)}| \tag{A.14}$$

$$= \frac{c_{ij}}{|C_{(i)}|}|C_{(l)}| \tag{A.15}$$

Therefore, the gradient of $\mathcal{F}_2(C)$ based on EqA.9:

$$\frac{\partial \mathcal{F}_2}{\partial c_{ij}} = \frac{\lambda_2}{2}\sum_{l=1}^{N} 2\cdot\omega_{il}\cdot\left(1-\frac{\langle C_{(i)},C_{(l)}\rangle}{|C_{(i)}||C_{(l)}|}\right)\cdot(-1)$$
$$\cdot\left(\frac{c_{lj}|C_{(i)}||C_{(l)}| - \langle C_{(i)},C_{(l)}\rangle\cdot\frac{c_{ij}}{|C_{(i)}|}|C_{(l)}|}{|C_{(i)}|^2|C_{(l)}|^2}\right) \tag{A.16}$$

$$= \lambda_2\sum_{l=1}^{N}\omega_{il}\left(1-\frac{\langle C_{(i)},C_{(l)}\rangle}{|C_{(i)}||C_{(l)}|}\right)$$
$$\cdot\left(\frac{\langle C_{(i)},C_{(l)}\rangle\cdot c_{ij}}{|C_{(i)}|^3|C_{(l)}|} - \frac{c_{lj}}{|C_{(i)}||C_{(l)}|}\right) \tag{A.17}$$

$$= \lambda_2\left[\sum_{l=1}^{N} w_{il}\left(1-\frac{\langle C_{(i)},C_{(l)}\rangle}{|C_{(i)}||C_{(l)}|}\right)\frac{\langle C_{(i)},C_{(l)}\rangle}{|C_{(i)}||C_{(l)}|}\right]\frac{c_{ij}}{|C_{(i)}|^2}$$
$$-\lambda_2\sum_{l=1}^{N} w_{il}\left(1-\frac{\langle C_{(i)},C_{(l)}\rangle}{|C_{(i)}||C_{(l)}|}\right)\frac{1}{|C_{(i)}||C_{(l)}|}c_{lj} \tag{A.18}$$

To have matrix formulation, we define the matrix $\mathbf{coC} = \left\{\frac{\langle \mathbf{C}_{(i)},\mathbf{C}_{(l)}\rangle}{|\mathbf{C}_{(i)}||\mathbf{C}_{(l)}|}\right\}$, the column vector $|\mathbf{C}|$ with entries being l_2 norms of rows of $\mathbf{C}$, and $|\mathbf{C}|^{-1}$ as its element-wise reciprocal. Further, define

$$\mathbf{W}_3 = \mathbf{W}\circ(\mathbf{1}_{N\times N} - \mathbf{coC})\circ\mathbf{coC}$$

and

$$\mathbf{W}_4 = \mathbf{W} \circ (\mathbf{1}_{N\times N} - \mathbf{coC}) \circ \left(|\mathbf{C}|^{-1}(|\mathbf{C}|^{-1})^\top\right),$$

with "$\circ$" being the Hadamard product of matrices. Consequently, the matrix form of the gradient of EqA.18 is

$$\nabla\mathcal{F}_2 = \lambda_2 \mathrm{diag}(\mathbf{W}_3\mathbf{1})\mathrm{diag}(|\mathbf{C}|^{-2})\mathbf{C} - \lambda_2\mathbf{W}_4\mathbf{C}. \tag{A.19}$$

A.3 Standard NMF multiplicative updates

We first derive the standard multiplicative update rules for NMF without the solvability and manifold penalties. Consider the reconstruction loss

$$S(\mathbf{C},\mathbf{P}) = \frac{1}{2}\|\mathbf{G} - \mathbf{CP}\|_F^2. \tag{A.20}$$

Equivalently, using the trace form,

$$S(\mathbf{C},\mathbf{P}) = \frac{1}{2}\,\mathrm{tr}\left[(\mathbf{G}-\mathbf{CP})^\top(\mathbf{G}-\mathbf{CP})\right]. \tag{A.21}$$

A.3.1 Update for $\mathbf{C}$

The gradient of $S(\mathbf{C},\mathbf{P})$ with respect to $\mathbf{C}$ is

$$\frac{\nabla S(C,P)}{\nabla C} = -(G-CP)P^\top = -GP^\top + CPP^\top \tag{A.22}$$

At iteration t, a gradient descent step for $\mathbf{C}$ can be written as

$$C^{(t+1)} = C^{(t)} - \alpha_c\left(-G(P^{(t)})^\top + C^{(t)}P^{(t)}(P^{(t)})^\top\right) = C^{(t)} + \alpha_c G(P^{(t)})^\top - \alpha_c C^{(t)}P^{(t)}(P^{(t)})^\top \tag{A.23}$$

To make sure the non-negativity of $C^{(t+1)}$, we let

$$\alpha_c = \frac{C^{(t)}}{C^{(t)}P^{(t)}(P^{(t)})^\top}$$

gives the multiplicative update

$$C^{(t+1)} = \frac{C^{(t)}}{C^{(t)}P^{(t)}(P^{(t)})^\top} \cdot G(P^{(t)})^\top = C^{(t)} \cdot \frac{G(P^{(t)})^\top}{C^{(t)}P^{(t)}(P^{(t)})^\top} \tag{A.24}$$

A.3.2 Update for $\mathbf{P}$

Similarly, the gradient of $S(\mathbf{C}, \mathbf{P})$ with respect to $\mathbf{P}$ is

$$\frac{\nabla S(C,P)}{\nabla P} = -C^\top(G - CP) = -C^\top G + C^\top CP \tag{A.25}$$

At iteration t, a gradient descent step for $\mathbf{P}$ is

$$P^{(t+1)} = P^{(t)} - \alpha_p \left(-(C^{(t)})^\top G + (C^{(t)})^\top C^{(t)} P^{(t)}\right) = P^{(t)} + \alpha_p (C^{(t)})^\top G - \alpha_p (C^{(t)})^\top C^{(t)} P^{(t)}. \tag{A.26}$$

To make sure the non-negativity of $P^{(k+1)}$, we let

$$\alpha_p = \frac{P^{(t)}}{(C^{(t)})^\top C^{(t)} P^{(t)}}$$

gives the multiplicative update

$$P^{(t+1)} = \frac{P^{(t)}}{(C^{(t)})^\top C^{(t)} P^{(t)}} \cdot (C^{(t)})^\top G = P^{(t)} \cdot \frac{(C^{(t)})^\top G}{(C^{(t)})^\top C^{(t)} P^{(t)}} \tag{A.27}$$

Thus, the standard NMF multiplicative updates are Eq. A.24 and Eq. A.27

A.4 GSNMF+ Multiplicative Updates

We next derive the multiplicative update rule for GSNMF+. Recall that the full objective function is

$$\mathcal{F}(\mathbf{C},\mathbf{P}) = \frac{1}{2}\|\mathbf{G}-\mathbf{CP}\|_F^2 + F_1(\mathbf{C}) + F_2(\mathbf{C}) + \mathbf{1}_T(\mathbf{P}). \tag{A.28}$$

From the previous sections, the gradients of the solvability and manifold penalties are

$$\nabla\mathcal{F}_1 = \lambda_1\mathbf{W}_1\left(\mathbf{W}_2\mathbf{C}-\mathbf{C}_g\right).$$

and

$$\nabla\mathcal{F}_2 = \lambda_2\mathrm{diag}(\mathbf{W}_3\mathbf{1})\mathrm{diag}(|\mathbf{C}|^{-2})\mathbf{C} - \lambda_2\mathbf{W}_4\mathbf{C}.$$

Since both $F_1(\mathbf{C})$ and $F_2(\mathbf{C})$ depend only on $\mathbf{C}$, the gradient of the full objective with respect to $\mathbf{C}$ is

$$\begin{aligned}
C^{(t+1)} &= C^{(t)} - \alpha_c\cdot\frac{\partial \mathrm{Obj}}{\partial C}\\
&= C^{(t)} - \alpha_c\cdot\left(-G(P^{(t)})^\top + C^{(t)}P^{(t)}(P^{(t)})^\top + \frac{\partial\mathcal{F}_1}{\partial c_{ij}} + \frac{\partial\mathcal{F}_2}{\partial c_{ij}}\right)\\
&= C^{(t)} - \alpha_c\Big[-G(P^{(t)})^\top + C^{(t)}P^{(t)}(P^{(t)})^\top + \lambda_1 W_1(W_2C^{(t)}-C_g) +\\
&\quad \lambda_2\mathrm{diag}(W_3\mathbf{1})\mathrm{diag}(|C^{(t)}|^{-2})C^{(t)} - \lambda_2 W_4C^{(t)}\Big]\\
&= C^{(t)} + \alpha_c\Big[G(P^{(t)})^\top + \lambda_1W_1C_g + \lambda_2W_4C^{(t)}\Big] - \alpha_c\left[C^{(t)}P^{(t)}(P^{(t)})^\top + \lambda_1W_1W_2C^{(t)} +\right.\\
&\quad \left.\lambda_2\mathrm{diag}(W_3\mathbf{1})\mathrm{diag}(|C^{(t)}|^{-2})C^{(t)}\right.
\end{aligned}$$

To ensure the non-negativity of $C^{(t+1)}$, we let

$$\alpha_c = \frac{C^{(t)}}{C^{(t)}P^{(t)}(P^{(t)})^\top + \lambda_1W_1W_2C^{(t)} + \lambda_2\mathrm{diag}(W_3\mathbf{1})\mathrm{diag}(|C^{(t)}|^{-2})C^{(t)}}$$

At iteration t, a gradient descent step for $\mathbf{C}$ can be written as

$$C^{(t+1)} = C^{(t)} \cdot \frac{G(P^{(t)})^\top + \lambda_1 W_1 C_g + \lambda_2 W_4 C^{(t)}}{C^{(t)} P^{(t)} (P^{(t)})^\top + \lambda_1 W_1 W_2 C^{(t)} + \lambda_2 \text{diag}(W_3 \mathbf{1}) \text{diag}(|C^{(t)}|^{-2}) C^{(t)}} \tag{A.29}$$

For P update,

$$P^{(t+1)} = P^{(t)} \cdot \frac{(C^{(t)})^\top G}{(C^{(t)})^\top C^{(t)} P^{(t)}} \tag{A.30}$$