Cytomulate: Accurate and Efficient Simulation of CyTOF data

Yuqiu Yang[1,2], Kaiwen Wang[1], Zeyu Lu[1], Tao Wang[2,3,*], Xinlei Wang[1,*]

1 Department of Statistical Science, Southern Methodist University, Dallas, TX, USA, 75275.
2 Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA, 75390.
3 Center for the Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, TX, USA, 75390.
* Corresponding authors

Short title: Cytomulate: Simulation of CyTOF data
Key words: CyTOF, simulation, proteomics

Abstract:

Cytometry by time-of-flight (CyTOF) is a proteomic analysis technique capable of providing information on millions of cells. Along with technological development, researchers have devised many data analysis tools to offer deeper insights into the data at the single-cell level. However, objective evaluations of these methods remain absent. In this paper, we develop Cytomulate: an accurate, efficient, and well-documented simulation model and algorithm for reproducible and accurate simulation of CyTOF data, which could serve as a foundation for future development and evaluation of new methods. With its two simulation modes: creation mode and emulation mode, Cytomulate is capable of generating simulated datasets that capture various characteristics of CyTOF data under different circumstances. We also provide the users with methods to simulate batch effects, temporal effects, and cell differentiation paths. Empirical experiments suggest that Cytomulate is superior in learning the overall data distribution than other single-cell RNA-seq-oriented methods such as Splatter.

Background:

Recent years have seen exponential growth in applications of mass cytometry (CyTOF) which characterizes proteomics profiles of tens of thousands or millions of heterogeneous single cells on 30-40 protein channels with minimal spill-over effects (1–3). As the technique of CyTOF matures, many data analysis methods have been proposed to offer insights into various aspects of

single-cell experiments (4). For example, in (5), the authors have observed that the signal intensity of a CyTOF experiment would vary over time due to changes in instrument performance. Their proposed method is currently the dominant tool to alleviate the temporal effect using polystyrene beads embedded with metal lanthanides. On the other hand, methods such as CytofRUV (6), and Cytonorm (7), etc. aim at getting rid of batch effects to facilitate comparisons across different CyTOF experiments. Deep learning algorithms such as residual networks (ResNet) (8) and generative adversarial networks (GAN) (9) were also adopted to resolve the same problem. While some software such as FlowSOM (10), LAMBDA (11), and Bayesian trees (12) focus on automatic cell type identification, others like CytoTree (13) attempt to explore the trajectories along which cells differentiate into other cells. Various data visualization and dimension reduction algorithms have also been applied to CyTOF (14). Typical examples include: t-SNE (15), UMAP (16), MDS (17), SAUCIE (18), scvis (19), etc. As the output of a CyTOF acquisition software is in Flow Cytometry Standard (FCS) format which might appear to be intimidating for researchers who are new to the field, packages such as flowCore (20) and Seurat (21) have incorporated standard methods to read, write, and preprocess CyTOF data.

However, most of the CyTOF analysis software developed by researchers is often evaluated against real data where the ground truth is typically unknown, rendering performance assessment subjective. Simulated datasets, albeit being synthetic, contain ground truth against which various methods can be tested. This makes simulation a vital tool in methodological development. There is a lack of formal simulation tools for CyTOF data, which serve as the foundation for developing and validating all other CyTOF data analysis tools. By contrast, various methods have been developed for simulating single cell RNA-sequencing (scRNA-seq) data, such as Simple, Lun, BASICS, and Splatter (22), among which Splatter is dominant.

Since both scRNA-seq and CyTOF are single-cell sequencing techniques, some data analysis tools originally developed in the context of scRAN-seq can be adapted to solve similar problems that rise in CyTOF. Here, we argue that directly applying scRNA-seq simulation tools such as Splatter to CyTOF simulation is not optimal. First of all, as one of the most commonly used data preprocessing steps, the inverse hyperbolic sine function with a cofactor of 5 is often used to normalize CyTOF data (23). Due to its nonlinearity, methods like Splatter that model untransformed expressions would inevitably yield biased results. What’s more, even on its original scale, the expression matrix returned by CyTOF often contains real-valued data or even negative values as opposed to integer-valued “counts” we encounter in scRNA-seq. To bridge this gap, we develop Cytomulate, the first formal simulation tool that is capable of accurately mimicking the behaviors of CyTOF data. Further, our framework combines the flexibility of generating complex, realistic data with efficiency and usability, allowing researchers to acquire data at scale and practitioners to easily simulate datasets for their workflows.

Results:

Characteristics of CyTOF data:

In this section, we will explore some of the prominent characteristics of CyTOF data to motivate the design choices of Cytomulate. The protein expressions associated with the first patient in the dataset Levine_32dim which is publicly available via the R package HDCytoData (24) will be used for demonstration.

High volume, high dimension, and heterogeneity: The expression matrix associated with a typical CyTOF experiment usually contains tens of thousands or even millions of cell events (rows), 30-40 protein channels (columns), along with 10-30 different cell types. For example, the data of the first patient in Levine_32dim has 72,463 cell events, 32 columns, and 14 cell types manually gated by the authors. Therefore, efficiency in both extracting major patterns from numerous cell events and generating a large dataset should be crucial for any potential simulation tools.

Zero-inflation: Like scRNA-seq data, the protein expressions observed in a CyTOF experiment often contain columns that are abundant with zeros or values close to zero. Biologically, these columns correspond to surface proteins that are often too scarce to be detected by the machine. Take, for example the CD4 T cells in the dataset, we see in Fig. 1a and Fig. 1b that the majority of the columns contain a high proportion of values close to zero.

Dependent columns: In Fig. 1c, we calculated the correlation matrix among all the protein channels using the data of CD4 T cells. It should be clear that markers such CD44 and CD45 are moderately correlated, which implies that assuming that the columns are independent might be too strong of an assumption. This suggests modeling the data jointly instead of marginally.

Normality: As one of the few most commonly used probability distributions that can be extended to higher dimensions with relative ease, Gaussian distribution is usually one of the first distributions to be considered in the presence of high dimensional continuous data. In Fig. 1d, we performed dimension reduction using Principal Component Analysis (PCA) to explore the normality of the data. If the data is Gaussian or near Gaussian, then the PCs will look somewhat elliptic. Using the expressions of CD4 T cells, we do not see a grotesque deviation, advocating for the usage of Gaussian distributions as basic building blocks.

Temporal effect: Although not present in Levine_32dim since it has already been preprossed, the signals of a single CyTOF experiment would typically vary over time due to changes in instrument performance as illustrated in (5). It appears that the signal drift (temporal effect), albeit varying from batch to batch, affects all cell types within one batch in a similar way.

Namely, the overall shape of the temporal effect may be shared by all cell types within a single batch.

Batch effect: Apart from the temporal effect, many researchers also reported the presence of batch effects across multiple CyTOF experiments due to sample collection, stimulation, staining, etc (25). In (7) they reported that not only are the batch effects cell-type specific, they also vary depending on protein channels.

Cell differentiation: Similar to scRNA-seq, a common task is to study cell differentiation where the expressions of an individual cell are assumed to vary according to some continuous path that connects one cell type to another. In the context of scRNA-seq, lots of methods have been developed to align cells on continuous paths with typical examples Slingshot and Monocle (26–28). To our knowledge, CytoTree (13) is the only available method of trajectory inference designed for CyTOF.

Cytomulate Framework

In light of the previous exploratory data analysis, we propose Cytomulate, an accurate and efficient simulation algorithm of CyTOF data based on Gaussian Mixture Models (GMM) (29). It can easily extract prominent characteristics from a large number of cell events commonly encountered in CyTOF experiments; it is capable of generating multiple realistic looking synthetic datasets within seconds; and it has a detailed documentation elucidating how all the components work together.

From a high-level overview, Cytomulate provides users with two simulation modes: creation mode and emulation mode to facilitate method development as well as to familiarize new researchers with properties of CyTOF data.

Creation mode aims at quickly generating synthetic datasets that capture some aforementioned characteristics of CyTOF data such as zero-inflation, column dependencies, and near normality. As this mode does not intend to extract the data distribution of some specific dataset, it is particularly useful for testing methods on a large scale. For example, in (14), they used this mode of Cytomulate to evaluate the quality of 20 dimension reduction methods from various aspects such as local and global structure preservation on 425 simulated datasets with different attributes like number of cell types and number of differentiation paths. This mode allows practitioners to precisely define the features of each simulated dataset with known ground truth, which is not often possible with real data. Further, it can also serve as a pedagogical tool for researchers who are not familiar with the format of a CyTOF dataset.

Emulation mode, on the other hand, is designed to learn the distribution of some real CyTOF data. Hence, it requires as inputs the expression matrix to mimic and the associated cell type labels. Although Cytomulate can potentially perform cell identification, which might involve some minor alterations and would probably yield a similar result to that of a k-means clustering, we made the design choice of including cell type labels as a mandatory input for the following two reasons: 1. According to (30) and (31) , k-means algorithm and GMM-based model (SWIFT) are often not the optimal choice for clustering CyTOF data; 2. There are already plenty of well-established cell identification algorithms and automatic gating techniques such as FlowSOM (10), Bayesian Tree (12), ACDC (32), etc, all of which have similar goals but slightly different design considerations. For example, using FlowSOM requires minimal prior knowledge on the cell population other than some crude idea on the number of cell types. On the other hand, Bayesian Tree and ACDC use a template as an input to facilitate the automatic gating process. Since we expect the users of the emulation mode to be somewhat experienced in the field, we think it would be better to let the users choose and apply their own cell identification algorithms that suit their situations the best. The resulting expression matrices of this mode would be more realistic compared with the ones generated by the creation mode.

To dive into the details, Cytomulate, at its core, uses a GMM with truncation below at 0 to approximate or generate the probability density of a given cell type. Although the exploratory data analysis in the previous section shows that a multivariate Gaussian might be a good candidate for modeling a specific cell type in a CyTOF experiment, this should be taken with a grain of salt since the clustering in Levine_32dim is quite detailed. However, we should not expect all users to be experts in identifying granular cell types. For example, there are 4 subtypes of B cells in Levine_32dim: mature B cells, plasma B cells, pre-B cells, and pro-B cells. If the users are only interested in B cells a single multivariate Gaussian would be inadequate in describing the data generating mechanism (see the comparison study in the next section for details). Therefore, we chose GMM for its flexibility and its mathematical tractability.

To cater for the needs of the development of more complex methods, in both simulation modes, users are allowed to vary cell abundances, noise levels, batch effects, temporal effects, as well as cell differentiation paths. The simulated batch effects, temporal effects, cell differentiation trajectories, as well as the noise are then added once we sample from the GMMs. In Fig. 1e, we outlined the framework of Cytomulate. Further details on the two simulation modes and the probabilistic model can be found in the Methods section.

We also modularized all the aforementioned functionalities so that each of them can be easily controlled or turned off. Cytomulate returns final expression matrices, the cell type labels associated with the expression matrices, as well as cell differentiation information for each cell event. In the following sections, we conduct some comparison studies and demonstrate the usage of Cytomulate through several simple examples.

Simulation Comparison Study

To the best of our knowledge, Cytomulate is the first formal simulation tool designed specifically for CyTOF experiments. Consequently, a comparison study among CyTOF simulation methods would not be feasible. Nevertheless, we still wish to demonstrate through a series of empirical experiments and examples the flexibility, the strengths, as well as some potential future improvements of Cytomulate.

The emulation mode is used throughout the comparison studies, where the following simulation models are included:

1. Splatter (22)
2. LAMBDA (11)
3. Cytomulate constrained to 1 component and diagonal covariance (Independent Gaussian)
4. Cytomulate constrained to 1 component and full covariance (Gaussian)
5. Cytomulate with full covariance and model selection via Bayesian Information Criterion (BIC) from GMMs with components varying from 1 to 9 (Cytomulate)

Here we give a brief description of Splatter and LAMBDA followed by the rationale behind the choice of the models to be compared.

Splatter, or specifically Splat, is originally designed for simulating single-cell RNA sequencing data. The model uses as a base a Gamma-Poisson mixture to simulate counts observed in scRNA-seq experiments. The authors also take into consideration outliers, library size, mean-variance trend, and technical dropout. As Splatter expects integers as inputs, to simulate CyTOF data which usually have been transformed into a numeric scale using the inverse hyperbolic sine function with a cofactor of 5, we can transpose the expression matrix, carry out its inverse transformation, round up the results to the nearest integer (23), apply Splatter to simulate counts, and transform the simulated expressions back into the numeric scale.

LAMBDA (Latent Allocation Model with Bayesian Data Analysis) contains a Bayesian hierarchical model designated for cell identification of CyTOF data. The model assumes that given a cell type, the cell expressions follow a zero-inflated multivariate Gaussian distribution. The cell types, on the other hand, are generated from a Categorical distribution whose parameters are controlled by some clinical information. Parameter estimation is accomplished by a stochastic EM algorithm.

We hope that the comparison between Splatter and the Independent Gaussian model would convince the readers that adapting a simulation tool designed for scRNA-seq data to CyTOF data is nontrivial and the end results might not be satisfactory. By taking into consideration the linear

dependencies among protein channels, the Gaussian model should show the value of joint modeling when compared with its independent counterpart. LAMBDA, due to its rigorous mathematical derivation, would provide us with a yardstick against which we can evaluate the quality of the simulated datasets in terms of the first and second moments. Finally, our core model, Cytomulate should demonstrate the merit of modeling non-normality.

To demonstrate the versatility of Cytomulate, five publicly available datasets from different species and diverse anatomic sites are collected. See Table 1 for details. All of these datasets have been bead-normalized and transformed by inverse hyperbolic sine function with a cofactor of 5. As datasets Levine_32dim and Samusik are not originally zero-inflated, we truncated the expressions below at 0 to ensure a fair comparison. Cell type information is also publicly available for the first 3 datasets. Cell types in Covid and LG were identified via FlowSOM in conjunction with manual gating by Dr. Wang at UTSW (14).

Cytomulate well approximates real data

In this comparison study, we will make use of all five datasets and include as candidate models Independent Gaussian, Gaussian, Cytomulate, LAMBDA, and Splatter. Three metrics are adopted to quantify the quality of the simulated datasets: 1. L2 distance between the simulated mean and the observed mean; 2. Frobenius distance between the simulated covariance and the observed covariance; 3. KL divergence between the simulated data and the observed data. See the Methods section for further details. For each dataset, the process is repeated 50 times. We will first show a detailed comparison using the dataset CyAnno.

In Fig. 2, we see that LAMBDA indeed gave the best estimation in terms of the mean with Cytomulate being a close second. We are surprised to see that Cytomulate outperformed LAMBDA in estimating the covariance structure. Further, Cytomulate outperforms all other methods benchmarked in this study with the smallest KL divergence. As the KL divergence is a measure of how similar two distributions are, we conclude that Cytomulate is able to yield a better approximation to the real data compared with other methods. This result also empirically demonstrated that improvements can be made if only the first two moments are considered. Curiously, Spatter, which is the industry standard for scRNA-seq data simulation, performs the worst in all benchmarks by an overwhelmingly wide margin, demonstrating that Splatter is suboptimal when simulating CyTOF data.

We further applied UMAP to exemplify the resulting expression matrices of each method (Fig. 3a - Fig. 3d, Supp. Fig. 1). Due to the granularity of the cell type labels included in the dataset, at the first glance, results returned by Cytomulate and LAMBDA resemble the real data. However, if we examine the blobs at the bottom right of the central cluster more closely, we see that Cytomulate depicts the data in far more detail than does LAMBDA. The preservation of small

clusters provides an advantage in preserving rare cell types and their unique phenotypes, whereas LAMBDA's resolution limitations tend to group these cells into adjoining clusters.

We now provide the rankings on the three metrics for all five datasets (Fig. 3e). As expected, by formally deriving a rigorous estimation procedure, LAMBDA is able to achieve the highest ranking in terms of mean and sometimes covariance. However, as we have seen previously, the sufficiency of only considering the first two moments highly depends on the data and the coarseness of the cell type labels. By using GMMs at the cell type level, Cytomulate always gives the best approximation to the real data as measured by the KL divergence. It should also be evident that Splatter always gives the worst performance due to the bias introduced by nonlinear transformations. Even working with Independent Gaussian would yield a performance boost.

Cytomulate is efficient

As a single CyTOF experiment could contain tens of thousands or millions of cell events along with 30 - 40 protein channels, computational efficiency including estimation and simulation could become a real issue. We conducted an analysis on the running time of each method by varying the number of cells and the number of markers using the CD8 Naive cells in the dataset CyAnno which contains 8,414 cell events and 39 protein channels. To be more specific, while evaluating the impact of sample size on estimation time and simulation time, we used all 39 protein channels and applied bootstrap sampling to generate expression matrices containing various numbers of cell events. We fixed the number of cell events and sampled the columns to be included when investigating the effect of the number of markers on processing time. We included Independent Gaussian, Gaussian, Cytomulate with 5-component GMMs (Restricted Cytomulate), Cytomulate, LAMBDA, and Splatter.

Overall both estimation time and simulation time appear to be linear with respect to the number of cells and the number of protein channels. In Fig. 4a, we see that for a moderate-size dataset containing 100,000, even Cytomulate with model selection from 9 models only takes around 2 minutes. If restrictions are imposed, the estimation can be done within seconds. On the other hand, LAMBDA, due to the usage of a stochastic EM algorithm for the model estimation, is twice as slow. It's also surprising to see that the number of markers has a large impact on the estimation time of LAMBDA compared with other methods (Fig. 4b). In terms of simulation time (Fig. 4c and Fig. 4d), we see that no matter what the model structure is, Cytomulate is always the most efficient under both circumstances. Splatter seems the least efficient when it comes to simulation.

Reflecting on the results, we think that Cytomulate, even with model selection, is capable of handling large datasets commonly encountered in CyTOF experiments.

Cytomulate is robust against upstream cell-typing

Although an array of cell type labels is a required input for the emulation mode, we do not expect high granularity of the labels. In this experiment, we demonstrate that even if the cell identification is not of high resolution, Cytomulate is still able to yield a satisfactory approximation of the data distribution. We included Independent Gaussian, Gaussian, Cytomulate, LAMBDA, and Splatter.

As we mentioned previously, there are 4 subtypes of B cells in Levine_32dim dataset: mature B cells, plasma B cells, pre-B cells, and pro-B cells. In the first setting, we provided every method with these detailed cell type labels. In the second setting, however, we amalgamated all the labels into B cells and fed this information to the methods.

We used UMAP to visually compare the results under two settings (Fig. 5). Although with the detailed labels, all methods except for Splatter seem to give acceptable approximations to the real data (Fig. 5 and Supp. Fig. 2), only Cytomulate is able to model the 4 major clusters and 1 small cluster present in the real data in the absence of detailed cell type information (Fig. 5b). For instance, if detailed information is not given, LAMBDA approximates the entire dataset with one Gaussian distribution which puts its mode around the middle of the two separated major clusters where the actual density is low (Fig. 5d).

To quantify the comparisons, we repeated the process 50 times and computed the Kullback-Leibler (KL) divergence (33) of the simulated datasets of each method under two settings (Fig. 5f). It should be obvious from the plot that although extra details would help improve the performance, Cytomulate appears to be the most robust method among all methods under comparison.

Contemplating on the results of this series of comparison studies, we see that Cytomulate is well suited for the large number of cell events commonly encountered in CyTOF experiments, is robust to upstream cell type identification results, and accurately captures the prominent characteristics of the data.

Complex simulations:

In this section, we will briefly demonstrate how the basic Cytomulate model used in the comparison study can be extended to incorporate more complex effects commonly encountered in CyTOF experiments such as batch effects, temporal effects, as well as cell differentiations.

Temporal effects:

As was noticed in (5), due to cell debris, etc. the sensitivity of CyTOF could potentially vary over time, causing the resulting signals to fluctuate even for the same cell type from one homogenous batch. In Cytomulate, we assume that the temporal effect for each batch is generated by a function defined on the unit interval such that the value at 0 always assumes 0, representing that there is no temporal effect at the beginning. We provide the users with 3 options to simulate a temporal effect, each with a different level of control over the overall shape of the temporal effect:

1. Brownian Bridge: The shape of the entire trajectory is randomly generated from a Brownian Bridge where the ending value is randomly sampled from a normal distribution. To evaluate the function on the unit interval, the Akima spline is used for interpolation. It should be clear that this option gives the users minimal control over the evolution of the temporal effects.
2. Polynomial: This option allows the user to specify the rough shape of the trajectory via the coefficients of a polynomial. A linear function will be used to make sure the resulting temporal effect starts at 0 and ends at a value randomly generated from a normal distribution.
3. Spline: During the preprocessing steps, it is possible that the users can extract the temporal effect component. The spline option permits the users to add the original temporal effect back to the simulated dataset via a smoothing spline. Since the size of the simulated dataset might differ from that of the original dataset, Cytomulate will automatically rescale the time interval so that the entire trajectory could be retained in the simulated dataset (Fig. 6a).

Batch effects:

As was noticed in (7), the batch effects in CyTOF experiments vary not only by batches but also by cell types and protein channels. To simulate batch effects, we decompose the effects into overall batch effects which assume different values for different batches and local batch effects which vary depending on the cell types and protein channels in a similar fashion to the way we decompose effects into main effects and interactions in Analysis of Variance (ANOVA). Further details can be found in the Methods section. As a demonstration, we simulated two batches with 3 cell types using the creation mode of Cytomulate (Fig. 8b).

Cell differentiation:

To simulate cell differentiations, Cytomulate adopts a two-step method. In the first step, a polytree (a directed acyclic graph whose underlying undirected graph is a tree) or a polyforest (a collection of nonoverlapping polytrees) is generated (34). In the Creation Mode, this is done by

first randomly grouping the cell types and applying the Kruskal's algorithm (35) with random weights to sketch the underlying tree or forest. In the Emulation Mode, this is accomplished by the Clauset-Newman-Moore greedy modularity maximization algorithm (36) (37) followed by the Kruskal's algorithm with the L2 distances among observed average cell means as weights. In both circumstances, to convert the graph to a directed acyclic graph (DAG), a depth-first search with a randomly selected node as the root on each graph component is performed. Inspired by Splatter (22), in the second step, we use a series of Brownian Bridges to generate the actual trajectories along the polytree or the polyforest. To illustrate, we used the creation mode of Cytomulate to generate 5 cell types aligned on one polytree (Fig. 6c).

Discussion and conclusions

In this paper, we introduced Cytomulate, a data analysis tool designed for simulating CyTOF data accurately and efficiently. It builds on Gaussian Mixture Models with a fast and robust estimation procedure and appealing mathematical properties. Further, Cytomulate features a well-defined probabilistic model to capture prominent aspects of CyTOF data while also incorporating two simulation modes: creation mode and emulation mode to cater to the needs of various researchers. In response to the poor usability of less popular tools in the field, our software package is not only well documented with detailed tutorials and guidelines hosted online (https://cytomulate.readthedocs.io) but also well maintained with clearly defined software support cycles, code assurance workflows, and an open source license. Our package is currently available on both PyPI and conda for installation via standard procedures.

To demonstrate its strength, we examined various aspects of Cytomulate in terms of estimation efficiency, simulation efficiency, tolerance towards coarse upstream cell type identification, and resemblance to real data. When compared with other potential competitors such as Splatter and and the core model of LAMBDA using five publicly available datasets, Cytomulate was able to achieve the highest ranking under most circumstances.Crucially, we found that Cytomulate is superior in approximating the overall data distributions regardless of the granularity of the input. Finally, we also illustrated the versatility of Cytomulate by simulating more complicated phenomena commonly encountered in a CyTOF experiment such as signal drifts (temporal effects), batch effects, and cell differentiations.

Simulation is a powerful, flexible and cost-effective tool for performance evaluation, which plays a vital role in methodological development. Analysis methods should work in scenarios for which they are designed and do whatever they claim to do (e.g., normalization, clustering, differential abundance testing, or better performance or improved computation efficiency over other methods). A common way to test this is to conduct simulation studies, where generating

multiple datasets, as many as we wish, with different parameters or assumptions, can be quickly done at least cost.

Currently, no formal method is available for accurately synthesizing CyTOF data. By contrast, various methods have been developed for simulating scRNA-seq data, among which Splatter is the dominant one. At the time Splatter was invented in 2017, the scRNA-tools database recorded about 120 tools designed for analyzing scRNA-seq data. By the end of 2021, this database recorded over 1100 tools, many of which have been tested and validated using simulated data. Evidently, analytical tools developed for CyTOF are far fewer than those for scRNA-seq. To bridge this enormous gap, there is an urgent need to carefully design a simulation model for CyTOF data, which can serve as a trustworthy test bed for development of new CyTOF analysis methods. We hope that researchers could greatly benefit from this introduction of Cytomulate when understanding and developing CyTOF data analysis tools.

Methods

Probabilistic model of Cytomulate:

The probabilistic model of Cytomulate is intended to model the □□□□□□□ $(./5)$-transformed CyTOF data. Suppose that we have $□ = 1, \cdots, □$ batches, $□ = 1, \cdots, □$ protein channels, $□ = 1, \cdots, □^{□}$ cell events for the □th batch, and $□ = 1, \cdots, □$ cell types. We denote the probability of observing the □th cell type in the □th batch by $\pi^{□}_{□}$. Then, for the □th cell event in the □th batch, we assume that the cell type is sampled from $\Pi^{□}_{□} \sim$ □□□□□□□□□□□□□$(□^{□})$, where $\pi = [□^{□}_{1}, \cdots, □^{□}_{□}]$ is a □-by-1 probability vector.

Suppose the cell type for the □th cell event in the □th batch is □. Then, we assume that, without cell differentiation or background noise, the latent expression is $□^{□}_{□} \sim □^{□}(□^{□}, □^{□})$, where $□^{□}(□^{□}, □^{□})$ denote some distribution with mean $\mu^{□}$, and variance-covariance matrix $\Sigma^{□}$. We use a Gaussian Mixture Model to approximate this distribution in Cytomulate.

The simulation of cell differentiation consists of two steps. Suppose cell type □ differentiates into $□^{□}$ "children" $□^{□}{}_{1}, \cdots, □^{□}{}_{□^{□}}$. Then, similar to the cell type generation, we assume that for the □th cell event whose cell type is □, the cell type it differentiates into is sampled from $\Omega^{□}_{□} \sim$ □□□□□□□□□□□□□$(□^{□})$, where $\omega^{□} = [□^{□}_{1}, \cdots, □^{□}_{□^{□}}]$ is a $□^{□}$-by-1 probability vector.

Given the realization for the □th cell event is $□^{□}$, we construct the differentiation trajectory $□^{□ \to □^{□}}: [0,1]^{□} \mapsto □^{□}$ whose components are sampled from a series of Brownian bridges such

that the $\square$th marker satisfies $\square_{\square}^{\square\to\square^{\square}}(0) = 0$, and $\square_{\square}^{\square\to\square^{\square}}(1) = \square_{\square}^{\square^{\square}} - \square_{\square}^{\square}$, where $\mu_{\square}^{\square}$ and $\mu_{\square}^{\square^{\square}}$ are the mean expression levels of the $\square$th marker of cell type $\square$ and cell type $\square^{\square}$, respectively. As $\square^{\square\to\square^{\square}}$ is a stochastic process, we denote its realization as $\square^{\square\to\square^{\square}}$. We further introduce an $\square \times 1$ random vector $\square_{\square}^{\square}$ composed of $\square$ independent random variables $\square_{\square,\square}^{\square} \sim \square\square\square\square(\square_{\square}^{\square,\square}, \square_{\square}^{\square,\square})$. During sampling, $\square_{\square}^{\square}$ will first be sampled and its realization $\square_{\square}^{\square}$ will be plugged into $\square^{\square\to\square^{\square}}$.

The batch effect to be added to the $\square$th cell event whose cell type is $\square$ in the $\square$th batch is $\Psi^{\square,\square} = \square(\square_{\square}^{\square} \succ 0) \circ (\square^{\square} 1 + \square^{\square,\square})$, where $\circ$ is the Hadamard product, $\square$ is an indicator function, $\succ$ is a component-wise inequality, $\psi^{\square} \sim \square(0, \square_{\square}^{2})$, with $\sum_{\square=1}^{\square} \square^{\square} = 0$, 1 is an $\square \times 1$ vector, $\theta_{\square}^{\square,\square} \sim \square(0, \square_{\square}^{2})$, with $\sum_{\square=1}^{\square} \square_{\square}^{\square,\square} = \sum_{\square=1}^{\square} \square_{\square}^{\square,\square} = 0$.
The temporal effect to be added to the $\square$th cell event in the $\square$th batch is $\Phi^{\square} = \square(\square_{\square}^{\square} \succ 0) \circ \square^{\square}(\frac{\square}{\square^{\square}})$, where $\square^{\square}(\frac{\square}{\square^{\square}})$ is the realization of a stochastic process $\square^{\square}: [0,1]^{\square} \mapsto \square^{\square}$, such that $\square^{\square}(0) = 0$.

With these setups, the simulated latent expression level for the $\square$th cell event whose cell type is $\square$ and who is differentiating into cell type $\square^{\square}$ in the $\square$th batch is found by $\square_{\square}^{\square} = \square_{\square}^{\square} + \square^{\square\to\square^{\square}}(\square_{\square}^{\square}) + \square^{\square,\square} + \square^{\square}$. The corresponding final expression level is $\square_{\square}^{\square} = \square(\square_{\square}^{\square} \succ 0) \circ \square_{\square}^{\square} + \square_{\square}^{\square}$, where $\square_{\square}^{\square} \sim \square$ and $\square$ is some error distribution.

Further details on the creation mode and the emulation mode

In this section, we provided some further details on the two simulation modes used in Cytomulate.

Creation mode: As no real data is required from the users, the initialization stage of the creation mode simply involves randomly generating average cell expressions as well as the GMMs to sample cell expressions from. If cell differentiation is to be simulated, the number of non-overlapping polytrees is needed as an extra input from the users. Cell types will then be randomly grouped and average cell expressions are randomly generated such that cells that are close to each other in a polytree will have similar expressions.

Emulation mode: The emulation mode works in a slightly different manner by exploiting the expression matrix and its associated cell type labels. Starting with the expressions of the $\square$th cell type where $\square = 1, \cdots, \square$, Cytomulate will fit a GMM to its corresponding expression matrix and use the resulting model as the model to sample cell expressions from. The Bayesian Information Criterion (BIC) is adopted for model selection. If simulating differentiation paths is sought after,

the averages of the observed cell expressions of all cell types are used to construct a polytree or a polyforest.

Metrics

In this section, we elaborate on the metrics we used in the comparison studies.
Suppose in the simulated batch we have $\square = 1, \cdots, \square$ protein channels, $\square$ cell events, and $\square = 1, \cdots, \square$ cell types. We further assume that the cell abundance for the $\square$th cell type is $\square_{\square} \geq 0$ such that $\sum_{\square=1}^{\square} \square_{\square} = \square$ and that by $\square_{\square}^{\square}$ is the simulated $1 \times \square$ cell expression vector for the $\square$th cell event of the $\square$th cell type. Then, for the $\square$th cell type, we calculate the simulated mean as $\square_{\square} = \frac{1}{\square_{\square}} \sum_{\square=1}^{\square_{\square}} \square_{\square}^{\square}$ and the simulated covariance as $\square_{\square} = \frac{1}{\square_{\square}-1} \sum_{\square=1}^{\square_{\square}} (\square_{\square}^{\square})^{\square} \square_{\square}^{\square}$. We then carry out the same procedure on the observed dataset and find the corresponding $\square_{\square}^{\square}$, and $\square_{\square}^{\square}$. The L2 distance between the simulated mean and the observed mean is found via $\square\square_{\square} = ||\square_{\square} - \square_{\square}^{\square}||_2$. The Frobenius distance (38) between the simulated covariance and the observed covariance is calculated as $\square\square_{\square} = ||\square_{\square} - \square_{\square}^{\square}||_{\square}$. And, the KL divergence $\square\square_{\square}$ is estimated following the method proposed in (33). Finally, the discrepancy measures for all the cell types are aggregated by a weighted average where the weights are decided by the cell abundances.

Bibliography

1. Lou X, Zhang G, Herrera I, Kinach R, Ornatsky O, Baranov V, et al. Polymer-based elemental tags for sensitive bioassays. Angew Chem Int Ed. 2007;46(32):6111–4.
2. Spitzer MH, Nolan GP. Mass cytometry: single cells, many features. Cell. 2016 May 5;165(4):780–91.
3. Bandura DR, Baranov VI, Ornatsky OI, Antonov A, Kinach R, Lou X, et al. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. Anal Chem. 2009 Aug 15;81(16):6813–22.
4. Chester C, Maecker HT. Algorithmic Tools for Mining High-Dimensional Cytometry Data. J Immunol. 2015 Aug 1;195(3):773–9.
5. Finck R, Simonds EF, Jager A, Krishnaswamy S, Sachs K, Fantl W, et al. Normalization of mass cytometry data with bead standards. Cytometry A. 2013 May;83(5):483–94.
6. Trussart M, Teh CE, Tan T, Leong L, Gray DH, Speed TP. Removing unwanted variation with CytofRUV to integrate multiple CyTOF datasets. eLife. 2020 Sep 7;9.
7. Van Gassen S, Gaudilliere B, Angst MS, Saeys Y, Aghaeepour N. Cytonorm: A normalization algorithm for cytometry data. Cytometry A. 2020 Mar;97(3):268–78.
8. Shaham U, Stanton KP, Zhao J, Li H, Raddassi K, Montgomery R, et al. Removal of batch effects

using distribution-matching residual networks. Bioinformatics. 2017 Aug 15;33(16):2539–46.

9. Upadhyay U, Jain A. Removal of Batch Effects using Generative Adversarial Networks. arXiv. 2019;

10. Van Gassen S, Callebaut B, Van Helden MJ, Lambrecht BN, Demeester P, Dhaene T, et al. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. Cytometry A. 2015 Jul;87(7):636–45.

11. Abe K, Minoura K, Maeda Y, Nishikawa H, Shimamura T. Model-based clustering for flow and mass cytometry data with clinical information. BMC Bioinformatics. 2020 Sep 17;21(Suppl 13):393.

12. Ji D, Nalisnick E, Qian Y, Scheuermann RH, Smyth P. Bayesian trees for automated cytometry data analysis. BioRxiv. 2018 Sep 19;

13. Dai Y, Xu A, Li J, Wu L, Yu S, Chen J, et al. CytoTree: an R/Bioconductor package for analysis and visualization of flow and mass cytometry data. BMC Bioinformatics. 2021 Mar 22;22(1):138.

14. Wang K, Yang Y, Wu F, Song B, Wang X, Wang T. A Comparison of Dimension Reduction Methods for Cytometry by Time-of-Flight Data. BioRxiv. 2022 Apr 27;

15. Laurens van der M, Geoffrey H. Visualizing Data using t-SNE. J Mach Learn Res. 2008;

16. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv. 2018;

17. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer; 2008.

18. Amodio M, Srinivasan K, van Dijk D, Mohsen H, Yim K, Muhle R, et al. Exploring Single-Cell Data with Multitasking Deep Neural Networks. BioRxiv. 2017 Dec 19;

19. Ding J, Condon A, Shah SP. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nat Commun. 2018 May 21;9(1):2002.

20. Hahne F, LeMeur N, Brinkman RR, Ellis B, Haaland P, Sarkar D, et al. flowCore: a Bioconductor package for high throughput flow cytometry. BMC Bioinformatics. 2009 Apr 9;10:106.

21. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun 24;184(13):3573-3587.e29.

22. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol. 2017 Sep 12;18(1):174.

23. Data scientist's primer to analysis of mass cytometry data [Internet]. [cited 2022 May 26]. Available from: https://biosurf.org/cytof_data_scientist.html

24. Lukas M. CS. HDCytoData. Bioconductor. 2018;

25. Leipold MD, Obermoser G, Fenwick C, Kleinstuber K, Rashidi N, McNevin JP, et al. Comparison of CyTOF assays across sites: Results of a six-center pilot study. J Immunol Methods. 2018 Feb;453:37–43.

26. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: cell lineage and

pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018 Jun 19;19(1):477.

27. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014 Apr;32(4):381–6.

28. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 2019 May;37(5):547–54.

29. Bishop CM. Pattern Recognition and Machine Learning (Information Science and Statistics). Softcover reprint of the original 1st ed. 2006. Springer; 2016.

30. Liu X, Song W, Wong BY, Zhang T, Yu S, Lin GN, et al. A comparison framework and guideline of clustering methods for mass cytometry data. Genome Biol. 2019 Dec 23;20(1):297.

31. Weber LM, Robinson MD. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. Cytometry A. 2016 Dec 19;89(12):1084–96.

32. Lee H-C, Kosoy R, Becker CE, Dudley JT, Kidd BA. Automated cell type discovery and classification through knowledge transfer. Bioinformatics. 2017 Jun 1;33(11):1689–95.

33. Perez-Cruz F. Kullback-Leibler divergence estimation of continuous distributions. 2008 IEEE International Symposium on Information Theory. IEEE; 2008. p. 1666–70.

34. Dasgupta S. Learning Polytrees. arXiv. 2013;

35. Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to Algorithms, 3rd Edition (The MIT Press). 3rd ed. Cambridge, Mass: The MIT Press; 2009.

36. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Phys Rev E. 2004 Dec 6;70(6).

37. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function using NetworkX. In: Varoquaux G, Vaught T, Millman J, editors. Proceedings of the 7th Python in Science conference. SciPy conference proceedings; 2008. p. 11–5.

38. Trefethen LN, David Bau III. Numerical Linear Algebra. 1st ed. Philadelphia: SIAM: Society for Industrial and Applied Mathematics; 1997.

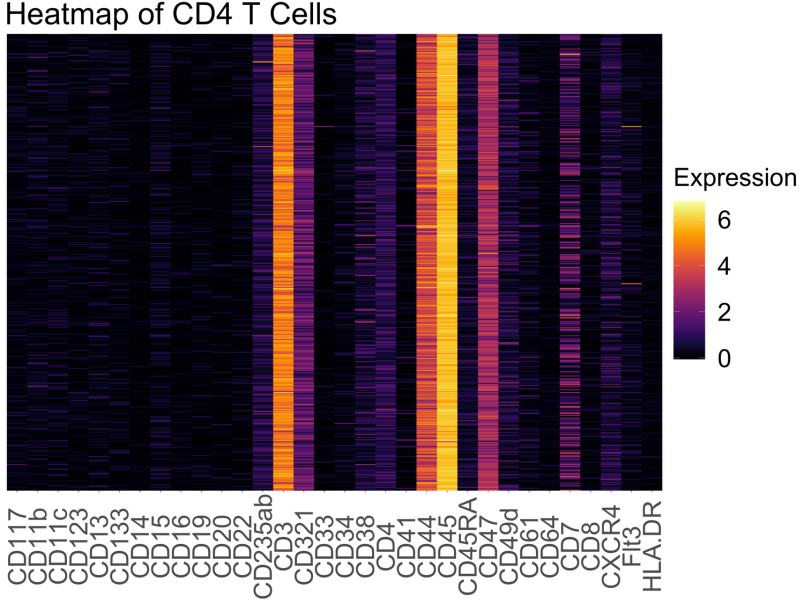

Fig. 1a

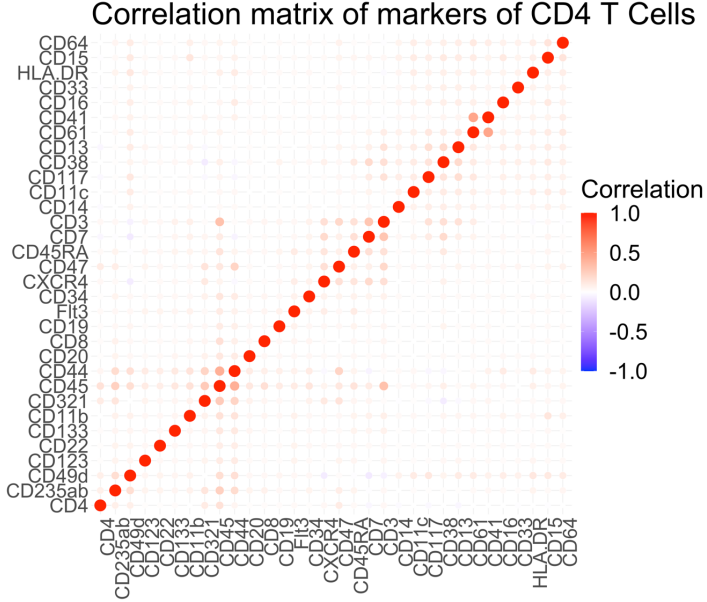

Fig. 1c

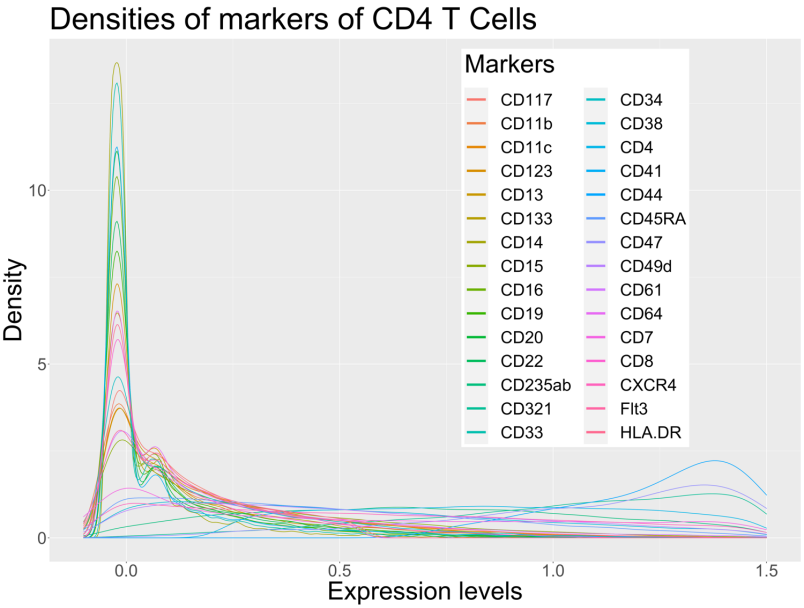

Fig. 1b

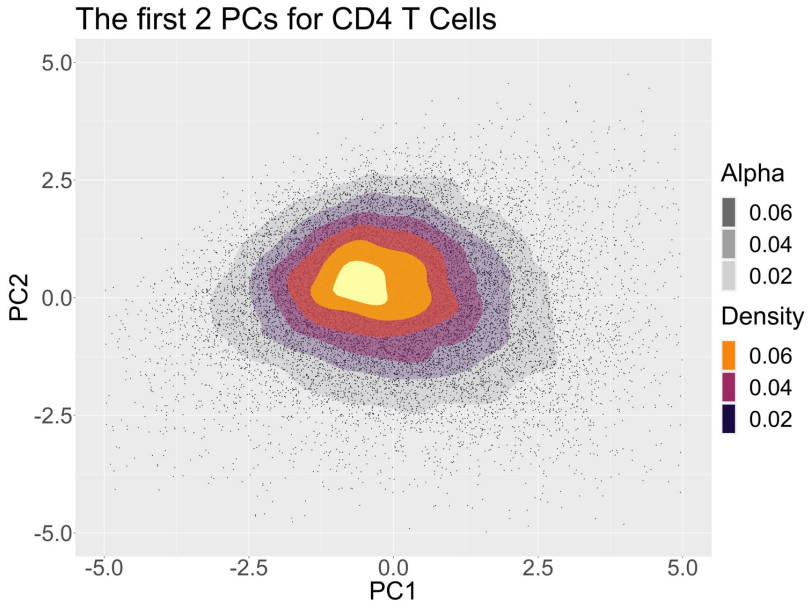

Fig. 1d

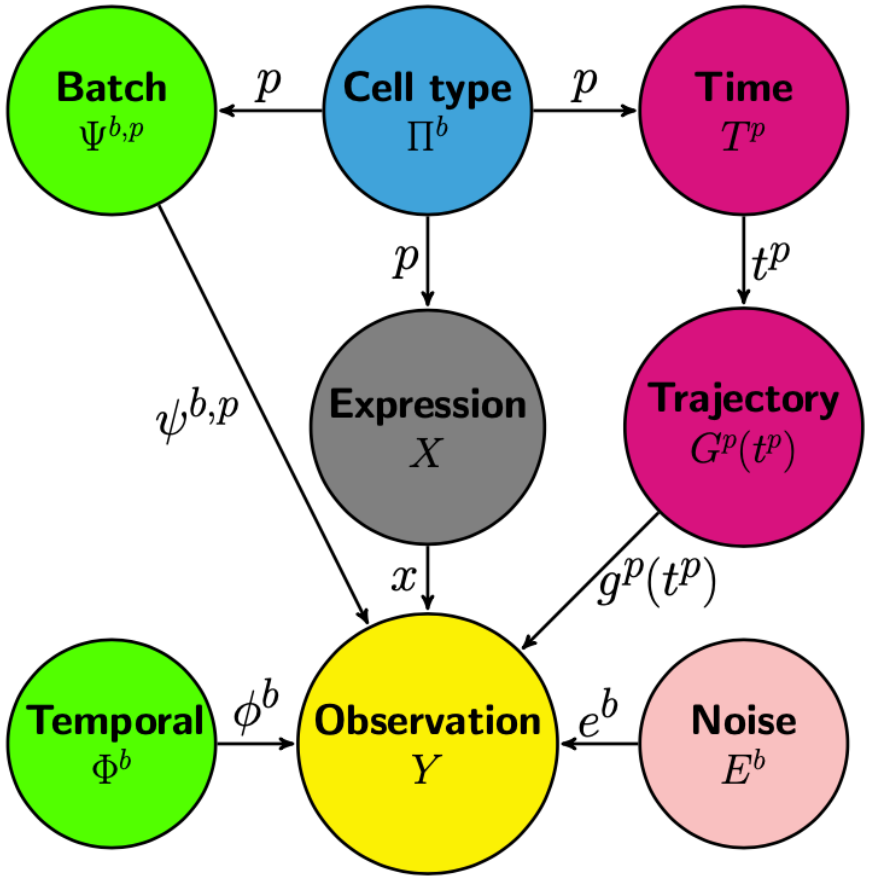

Fig. 1e

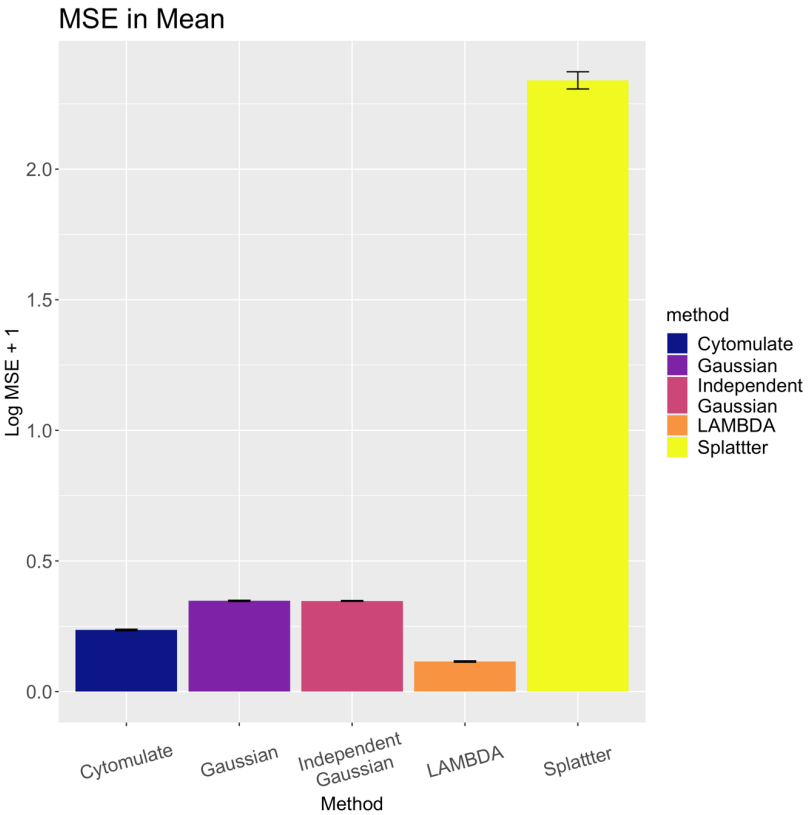

Fig. 2a

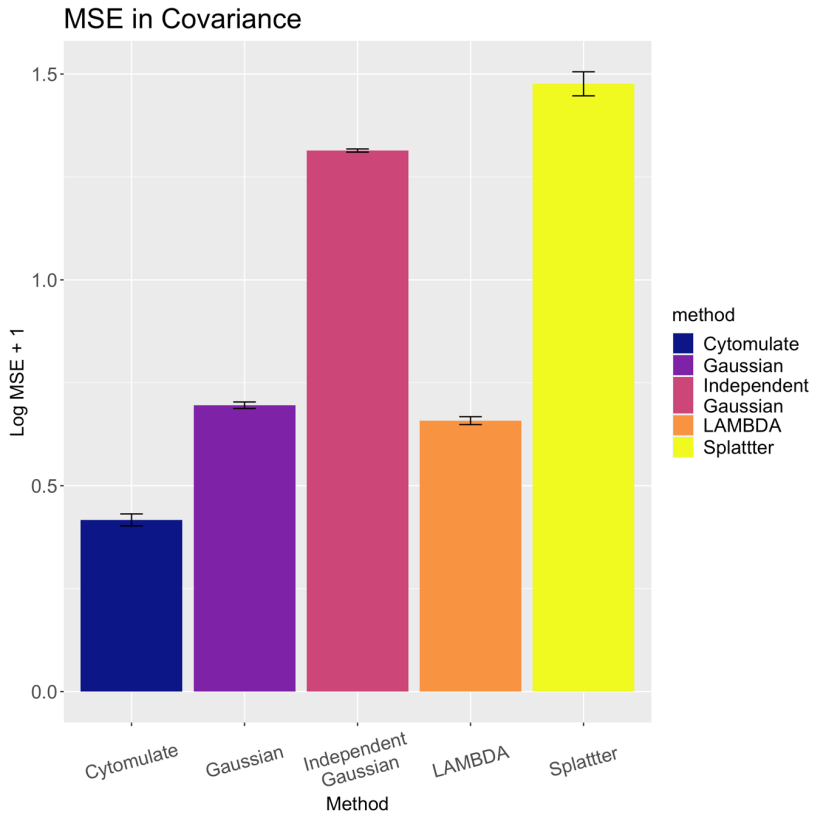

Fig. 2b

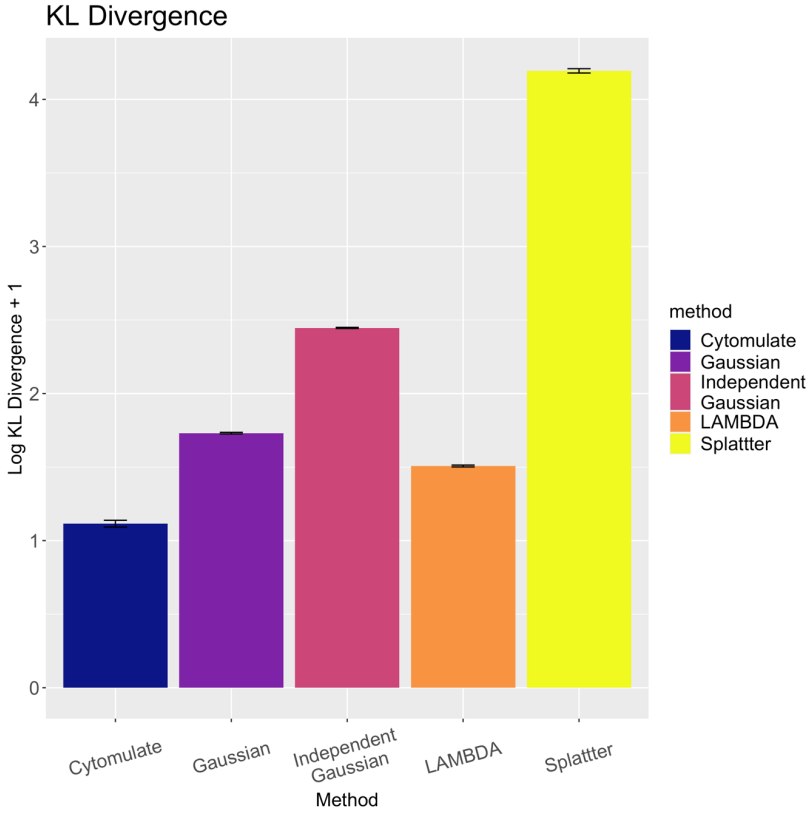

Fig. 2c

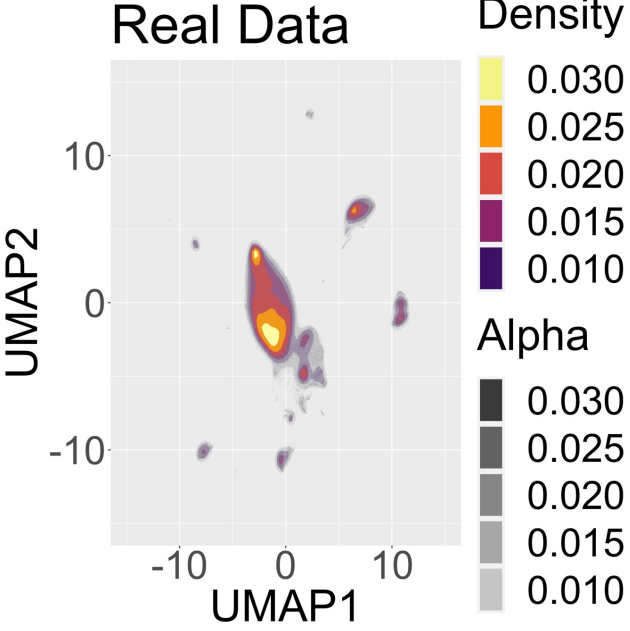

Fig. 3a

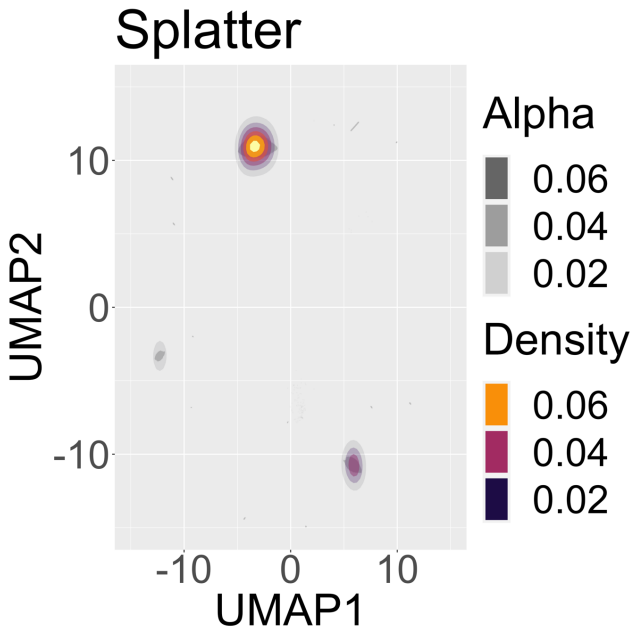

Fig. 3b

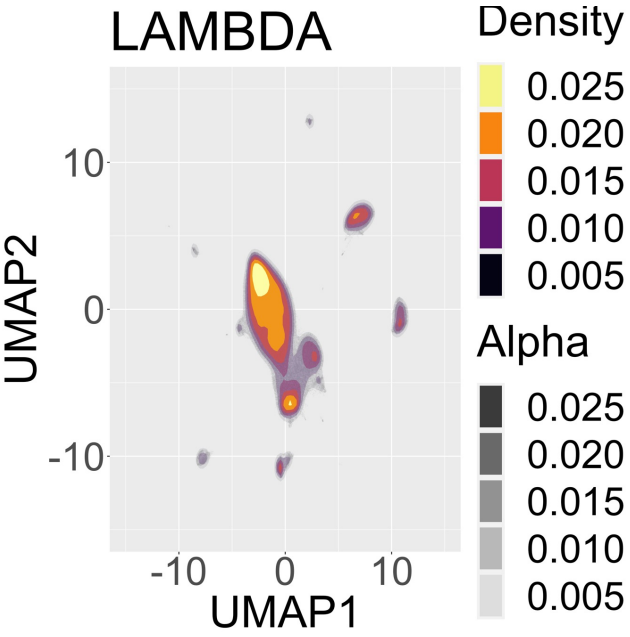

Fig. 3c

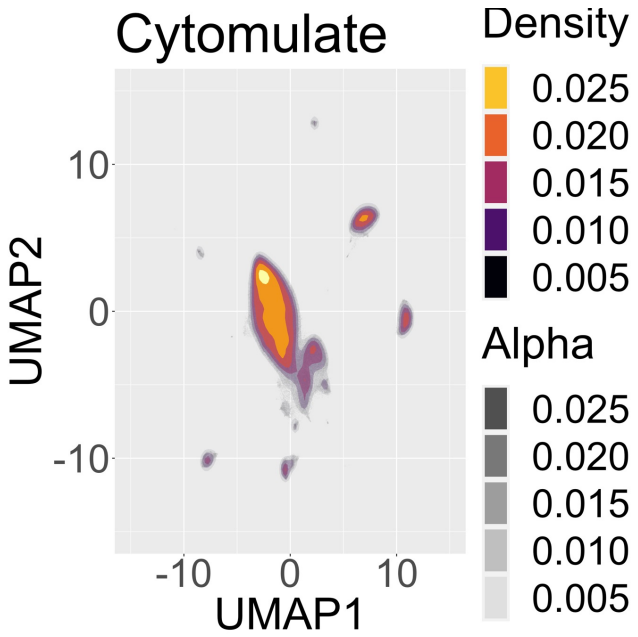

Fig. 3d

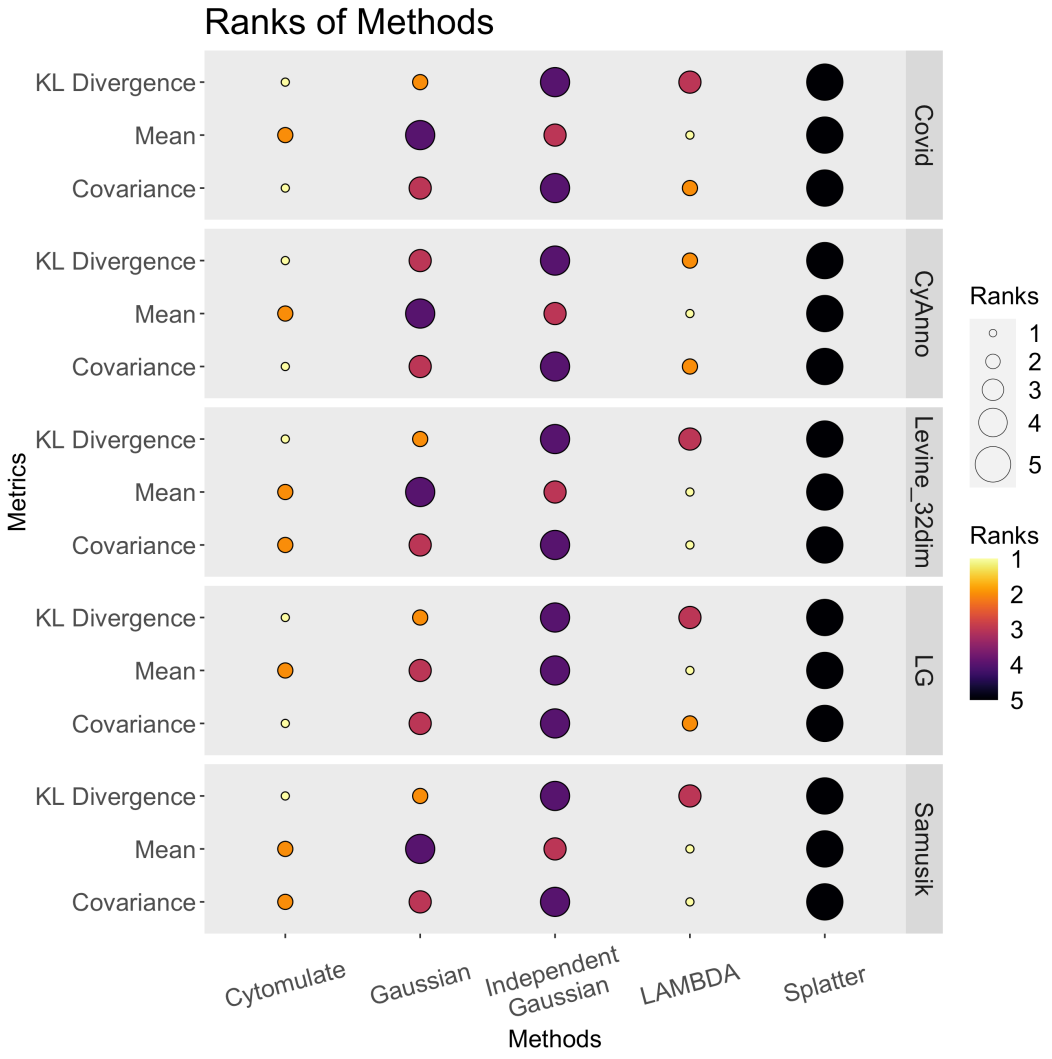

Fig. 3e

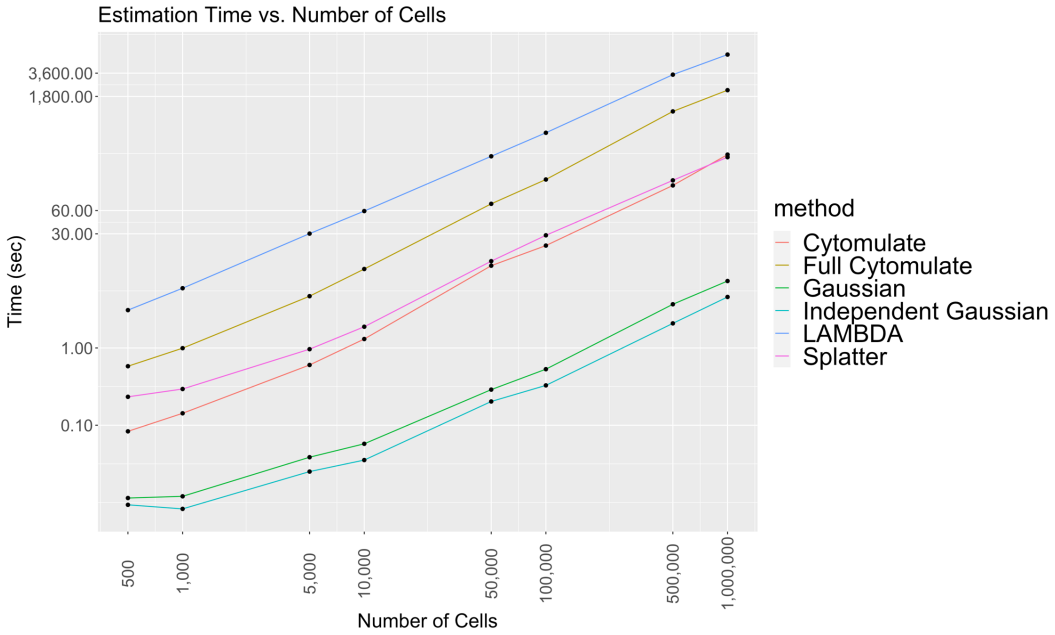

Fig. 4a

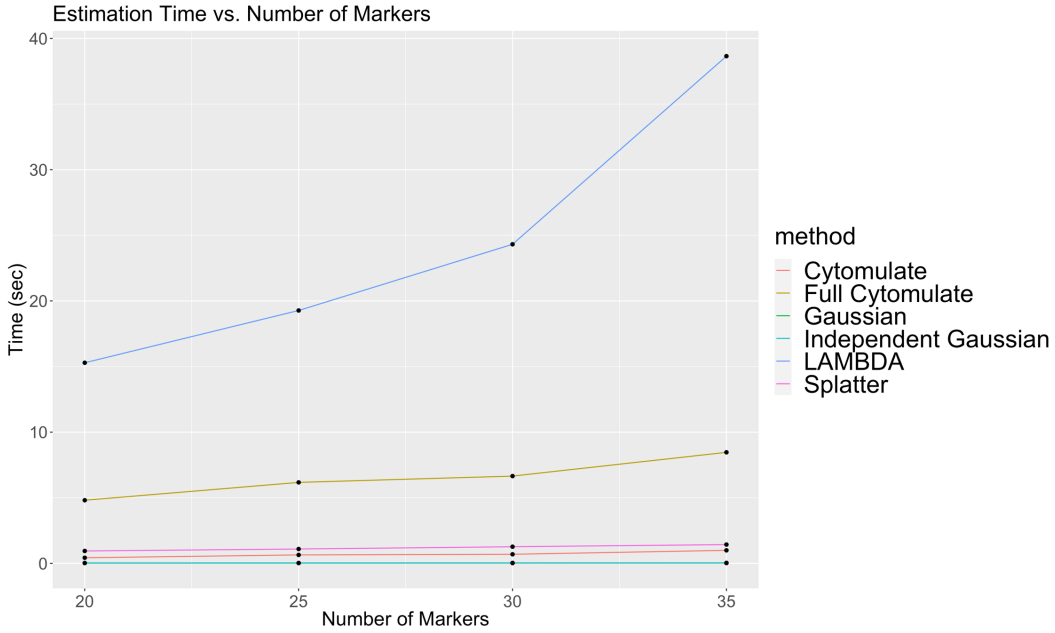

Fig. 4b

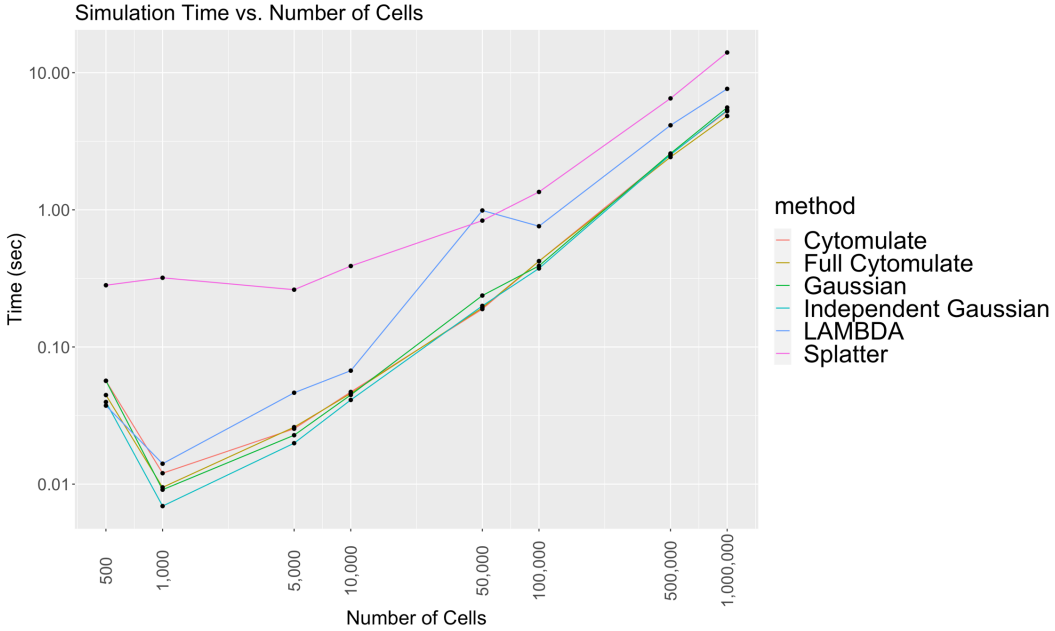

Fig. 4c

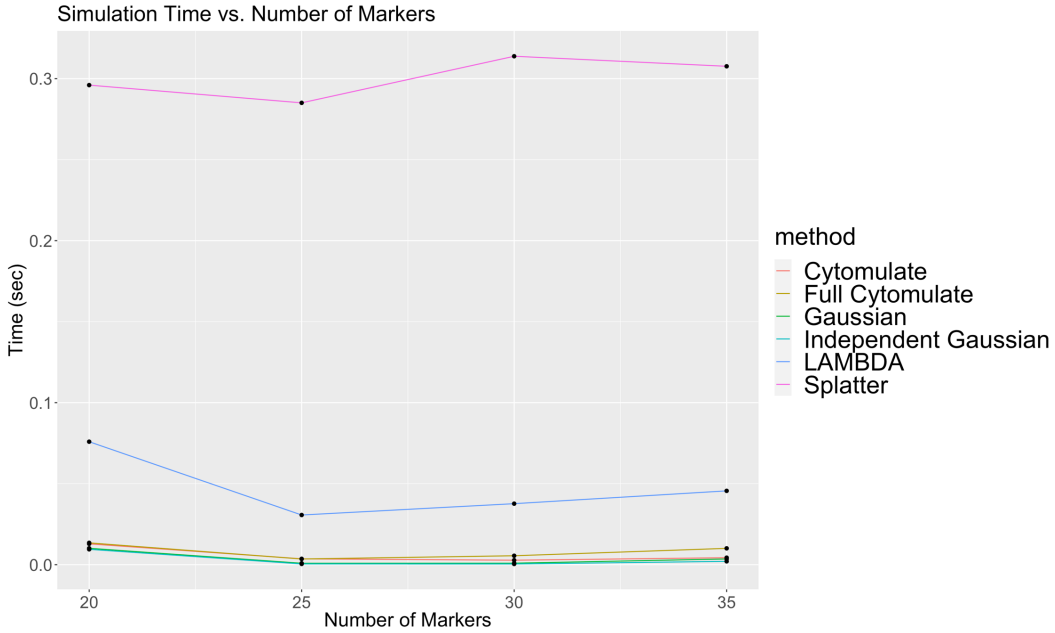

Fig. 4d

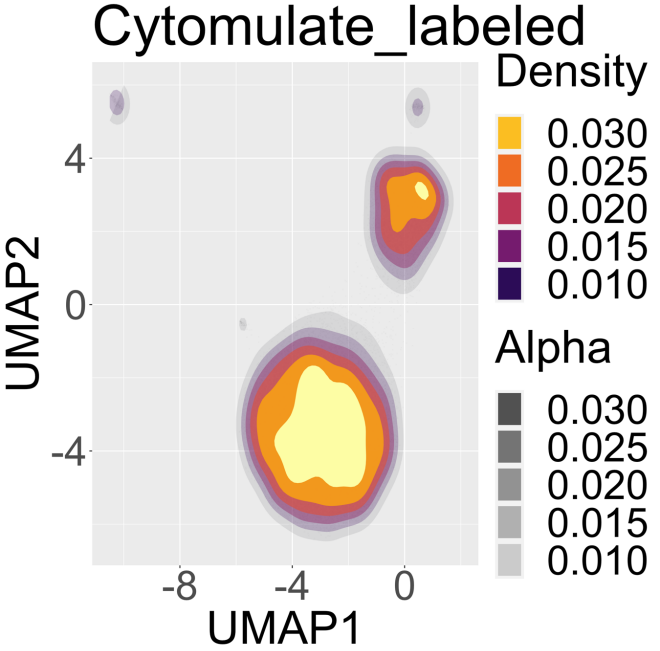

Fig. 5a

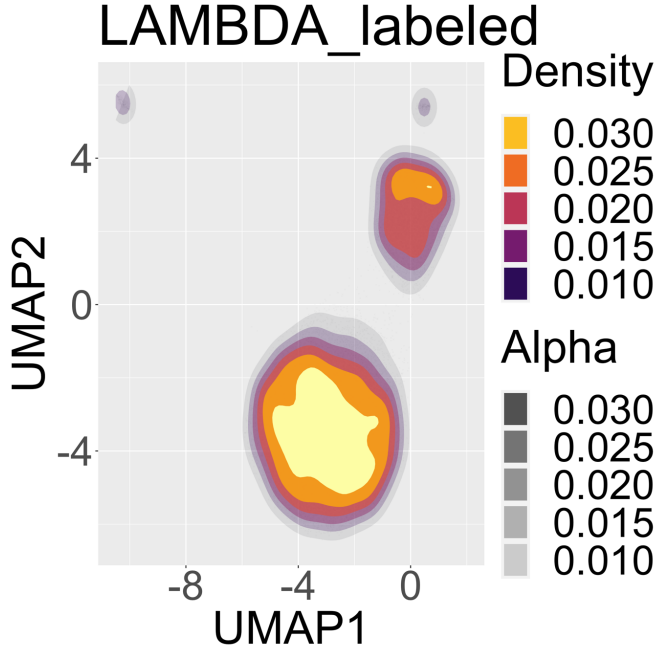

Fig. 5c

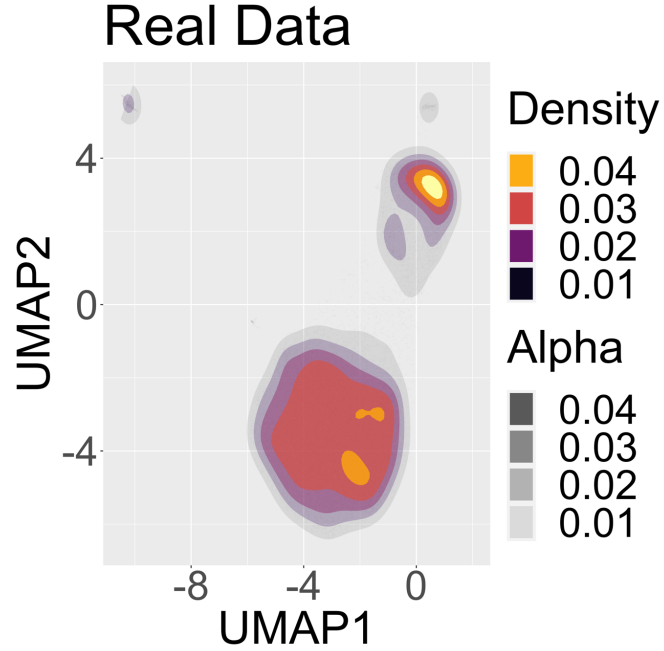

Fig. 5e

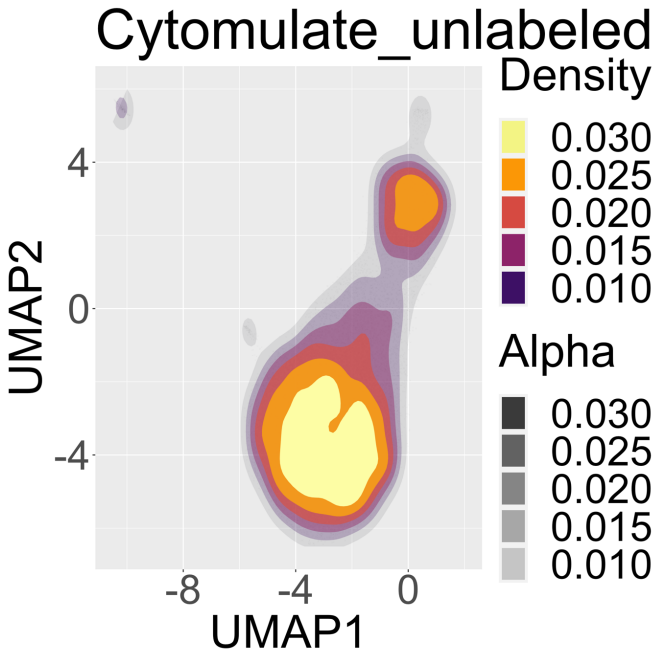

Fig. 5b

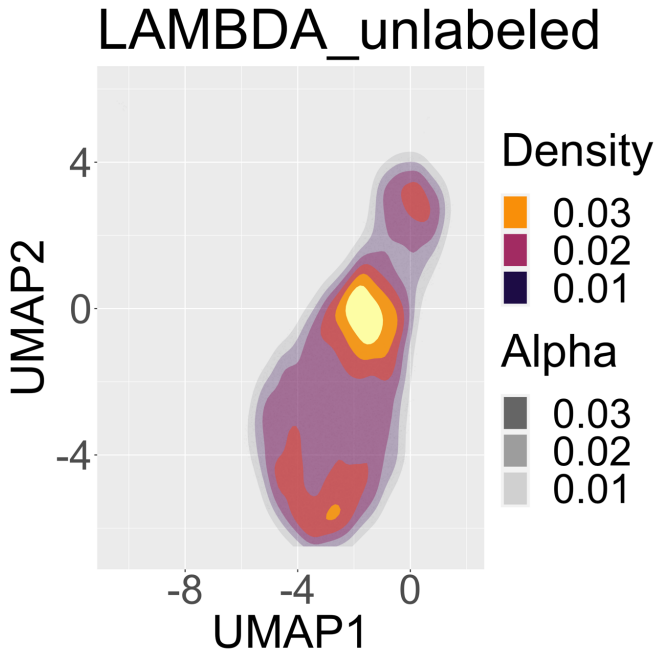

Fig. 5d

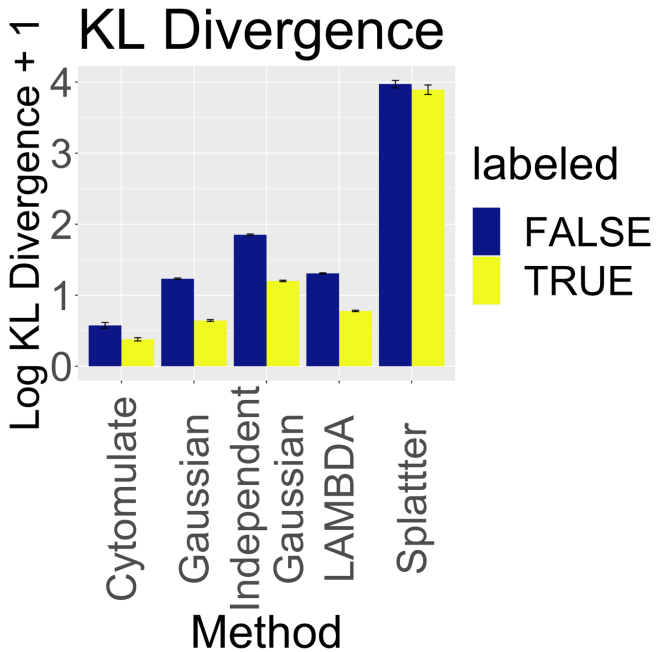

Fig. 5f

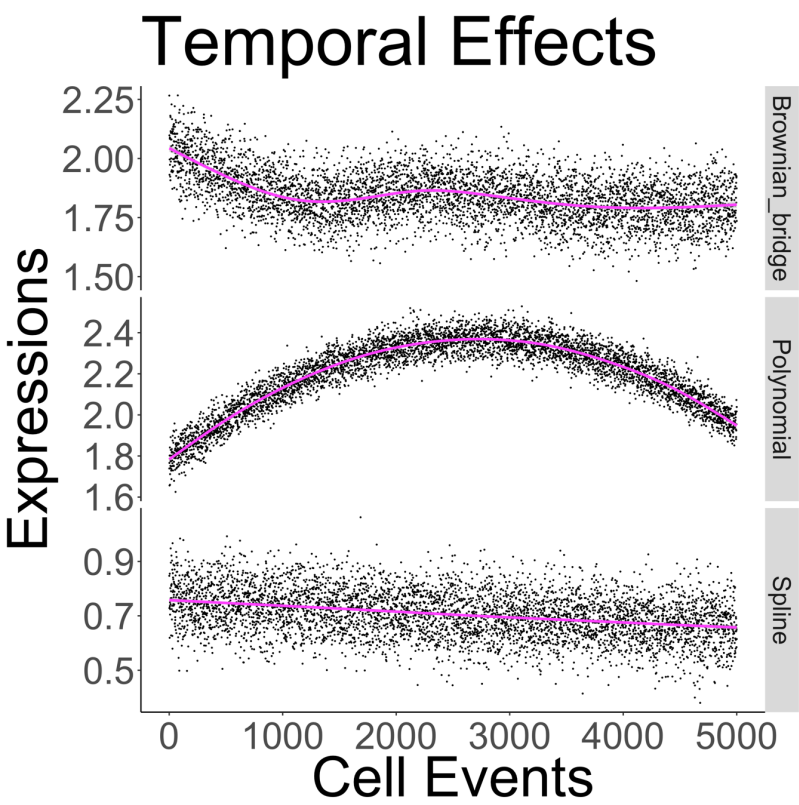

Fig. 6a

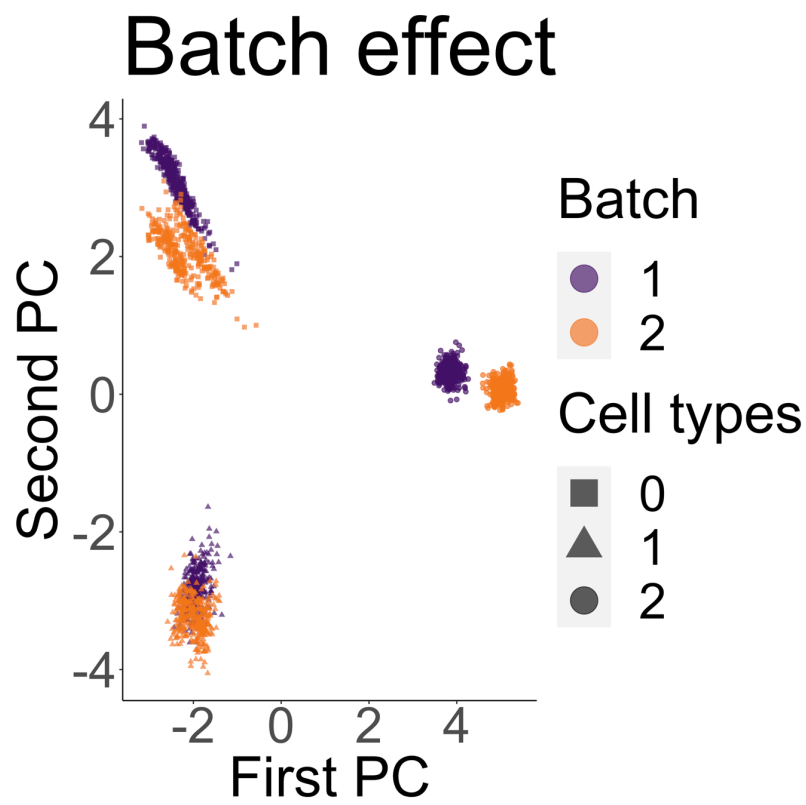

Fig. 6b

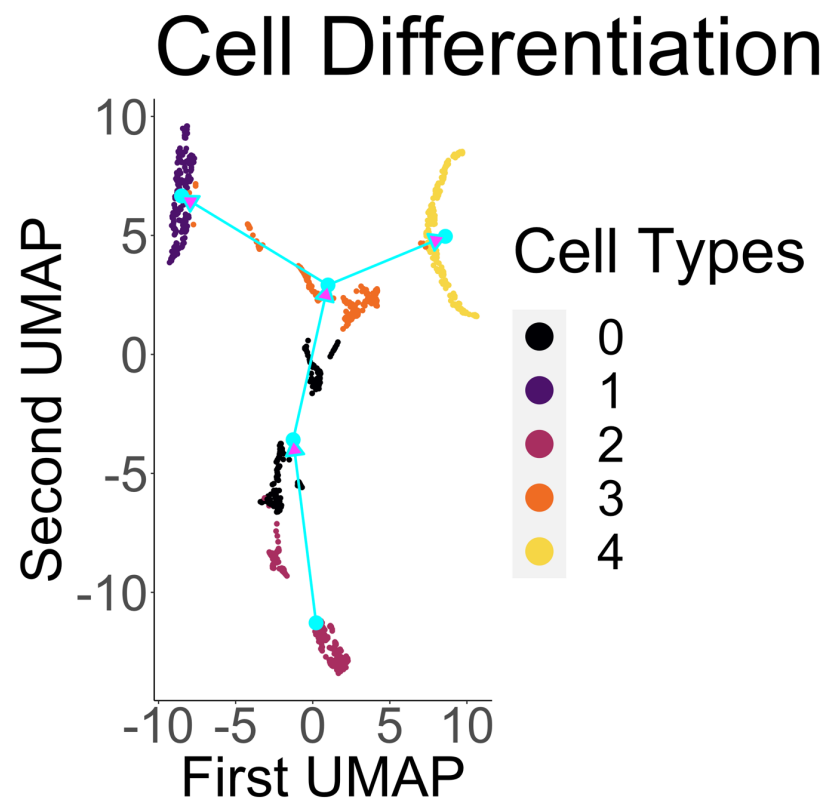

Fig. 6c

Dataset	Species	Anatomic Site	Source	Number of Markers
Levine_32dim	Human	Bone Marrow	HDCytoData	32
Samusik	Mouse	Bone Marrow	HDCytoData	39
CyAnno	Human	Peripheral Blood	https://flowrepositor	39
Covid	Human	Peripheral Blood	https://dbai.biohpc.s	43
LG	Mouse	Lacrimal Gland	https://flowrepositor	31

Number of Cells Events	Number of Cell Types
72,463	32
53,173	24
123,033	39
208,712	21
284,055	7