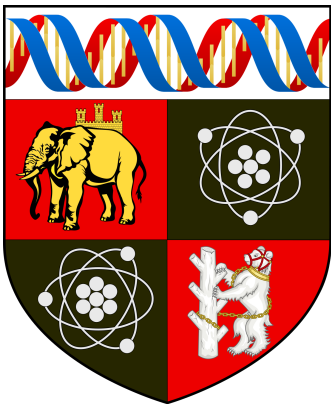


Scalable Graph Based Single Cell Omics Analysis

by

Shobana Venkat Stassen

Covering Document

Submitted to the University of Warwick
for the degree of
Doctor of Philosophy

School of Life Sciences

October 2022

List of Figures

Acknowledgements

The journey towards a PhD was only possible with the help of family, colleagues and supervisors spanning two continents. First and foremost, thank you to my husband and children for their patience and encouragement during various stages of research and manuscript preparation, not least the grueling hours spent during revisions and bouncing back from reviewer feedback. Thank you to my parents for nurturing and instilling in me a deep respect and admiration for researchers patiently pushing the frontiers of science. Secondly, thank you to Professor Kevin Tsia at Hong Kong University's ALPHA Laboratory for his kind, patient and thoughtful approach to research, for allowing me the freedom and independence to pursue areas in single cell omics analysis that were of interest to me, and providing the faith and encouragement that I could continue my research even after relocating to Singapore a few years ago - Work From Home became the norm for me well before the rest of the world switched to do the same! My intentions to stay in academic research can surely be attributed to the example set by Professor Tsia over the past years. Finally, thank you to my supervisor Daniel Hebenstreit at the University of Warwick, who took me in when the world locked down due to Covid and I could not travel to Hong Kong to complete my PhD due to the city's strict travel bans. Dr Hebenstreit's commitment and willingness to help me use my publications and research towards a degree have lifted my prospects for a career in academic research.

Declarations

This covering document is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work presented in this thesis was carried out by the author in collaboration with Professor Kevin Tsia (Alpha Lab in the University of Hong Kong). Conceptualization and planning was carried out by Shobana V. Stassen and Kevin Tsia. All novel computational modeling, computational design and analysis mentioned in this covering document was carried out by the author (Shobana V. Stassen). The biological data used for testing and analysis was publicly available and is appropriately referenced, or in the case of imaging cytometry data (Chapter 4) was generated by members of the Alpha Lab HKU and has been credited authorship in the relevant publications highlighted in this covering document.

The work presented in Chapter 4 uses data from imaging cytometry datasets. These datasets (lung cancer cells and cell cycle cells) are from experiments designed and carried out by Dr. Dickson M. D. Siu and Dr. Gwinky. G. K. Yip from the HKU Applied Life Photonics Laboratory. Michelle Lo, Dickson M.D. Siu and Gwinky G.K. Yip performed the image processing on these cells. Michelle Lo and Shobana V. Stassen collaborated on the transfer learning based analysis in Chapter 4.2.2.

This declaration is consistent with the authorship declarations made in the relevant publications and have been signed by the above-mentioned collaborators in a letter submitted together with this document.

Abstract

The last few years have seen tremendous growth in the generation of large scale, high dimensional complex single cell datasets that map cellular heterogeneity and development across entire organisms. The analysis of these ‘cellular atlases’ in order to harness useful biological insights into the development of healthy tissues and organs, as well as pathogenesis, places new requirements on the capabilities of single-cell analysis computational tools. This covering document summarizes the key contributions of the author towards two single-cell analysis tasks that are common to many single-cell pipelines, namely clustering [PARC Stassen et al., 2020] and trajectory inference [VIA Stassen et al., 2021]. The complexity and stochastic nature of single-cell data presents various challenges to its analysis. Certain distortions and computational bottlenecks only manifest themselves at high cell counts or high dimensionality. Many recent efforts are aimed at efficiently distilling information without resorting to techniques that oversimplify the data in terms of cell count (e.g. subsampling which may remove rarer cells and reduce heterogeneity) or excessive dimensionality reduction (relying on just a few dimensions from an embedding that loses global neighborhood relationships). The two new methods introduced in the text both utilize graph-based approaches to modeling single cell data, with an emphasis on maintaining accuracy in terms of detecting cell types, preserving inter-cellular relationships, and offering a fast and data driven approach to parameter selection even as the scale of cells exceeds 100,000s and even million of cells. The performance of PARC and VIA have been validated on a wide range of datasets and the methods have been well received by the single-cell community as seen by the download statistics and integration into various pipelines.

Published work and conference presentations indicating contribution and impact

Publications:

Shobana V Stassen, Yip, G.G.K., Wong, K.K.Y. et al. Generalized and scalable trajectory inference in single-cell omics data with VIA. Nat Commun 12, 5528 (2021). https://doi.org/10.1038/s41467-021-25773-3 **Contribution** (as reflected in the Contributions section of the published manuscript)**:** First author. Conceptualized and implemented the algorithm and interface for the VIA trajectory inference method. Computationally processed, analyzed and benchmarked public and in-house data. Wrote the main manuscript and supplementary information. All datasets except are/have been made publicly available. The imaging cytometry data of cell cycle progression was generated by Gwiky.GY.Yip. **Impact: Altmetric Score on Bioarxiv 51, Altmetric Score on Nature Communications 39 (Top 5% of research ranked on altmetric)**. **Citations**: **7**

Shobana V. Stassen, Dickson M. D. Siu, Kelvin C. M. Lee, Joshua W. K. Ho, Hayden K. H. So, Kevin K. Tsia. "PARC: ultrafast and accurate clustering of phenotypic data of millions of single cells". Bioinformatics. 36(9): 2778-2786 (2020). **Altmetric Score 50 (Top 5% of research). Contribution:** First author. Conceptualized and implemented the algorithm and interface for the PARC clustering method. Computationally processed, analyzed and benchmarked public and in-house data. Wrote the main manuscript and supplementary information. The lung cancer imaging flow cytometry data was generated by Dickson MD Siu. **Impact: Altmetric Score 49 (Top 5% of research).** Integrated into multiple single-cell data processing pipelines including those at 10X Genomics **Citations: 33**

Siu, Dickson, Lee Kelvin, Lo Michelle, **Shobana V. Stassen,** Wang Maolin, Zhang Iris, So Hayden, Chan Godfrey, Cheah Kathryn, Wong Kenneth, Hsin Michael, Ho James, Tsia Kevin. "Deep-learning-assisted biophysical imaging cytometry at massive throughput delineates cell population heterogeneity." Lab on a chip. (2020) **Contribution:** Implemented (Coded and designed) the transfer learning protocol for the neural network used for the single cell image-based data in the paper. Conducted preliminary testing with initial datasets to test and tune the hyperparameters. **Citations: 17. Impact: Altmetric 6. Cover page article Nov 2020**

Conferences

Shobana V Stassen, Kelvin C. M. Lee, Kevin K. Tsia, "Accelerated Pheno-Tree (APT) for large-scale, label-free of image-based single-cell analysis". High-Speed Biomedical Imaging and Spectroscopy IV 10889, SPIE Photonics West BIOS Speaker (Feb 2019). **Contribution:** Conceptualized and implemented the algorithm and interface for the APT clustering method. Computationally processed, analyzed and benchmarked public and in-house data. Wrote and delivered the presentation.

Michelle C.K. Lo, **Shobana V Stassen,** Dickson M.D. Siu, and Kevin K. Tsia, "Robust Quantitative Phase Imaging Cytometry with Transfer Learning". Biophotonics Congress: Biomedical Optics 2020 (Translational, Microscopy, OCT, OTS, BRAIN). **Contribution:** Implemented (Coded and designed) the transfer learning protocol for the deep neural network used for the single cell image-based data in the presentation. Conducted preliminary testing with initial datasets to test and tune the hyperparameters.

Shobana V Stassen, K. K. Tsia, "VIA for generalized and scalable single-cell trajectory inference beyond transcriptomic data". CYTO Virtual Interactive 2021 Speaker. **Contribution:** Conceptualized and implemented the algorithm and interface for the VIA method. Computationally processed, analyzed and benchmarked various in-house data. Wrote and delivered the presentation.

Gwinky Yip, Alex Chin, **Shobana V Stassen,** Michelle CK Lo, Rashmi Sreeramachandramurthy, Kelvin CM Lee, Kenneth KY Wong, Leo LM Poon, Kevin K Tsia, Image-based single-cell biophysical phenotyping of SARS-CoV-2 infection by high-throughput quantitative phase imaging flow cytometry, High-Speed Biomedical Imaging and Spectroscopy VII SPIE 2022. **Contribution:** Developed the trajectory inference model used on the data comprising Covid infected cells measured at 3 distinct time intervals after infection. Ran preliminary computational tests on the data.

Overview of covering document

The covering document provides the context and motivation behind the published materials. It summarizes the key algorithmic contributions and their impact in terms of advancing computational approaches to single cell omics datasets.

Chapter 1 is an overview of the field of single cell omics analysis. In particular, it describes the paradigm shift occurring in computational approaches to probing single cells as stochastic units as opposed to the averaged bulk measurements of the previous decade. This section describes computational steps present in single-cell omics analysis pipelines spanning data pre-processing (quality control, normalization, dimensionality reduction, batch correction), data visualization, clustering, classification, trajectory inference and differential analysis. Subsequent chapters address specific elements of this pipeline to which the published materials contribute.

In Chapter 2 we introduce a new clustering method PARC [Stassen et al., 2019] that uses hierarchical graph pruning combined with community detection to analyze cellular heterogeneity and detect rare cell populations. Clustering is typically used to capture the heterogeneity presented by single cell data in a concise representation of discrete groups that captures natural similarities. This enables researchers to then identify cells in terms of their ontology. In addition to aiding cell annotation, clustering can also be an intermediate step before other downstream analysis such as trajectory inference (as we show in Chapter 3) or differential/compositional analysis. We show that PARC addresses issues related to scalability (in cell count and dimensionality) and captures rare populations in a variety of single cell data.

Chapter 3 focuses on computational methods for the analysis of data capturing continuous differentiation processes such as those in development and regeneration. The prediction of cell ordering and branching processes is known as Trajectory Inference (TI). TI sheds light on the temporal dynamics and choices made at the single-cell level as cells transition from pluripotent states towards increasingly specialized ones that make up the heterogeneity seen in tissue and organs. We present a new graph based probabilistic approach called VIA [Stassen 2021] that projects the stochastic information from a single-cell level onto a cluster level graph in order to make predictions about the differentiation topology. We rely on our clustering method PARC (from Chapter 2) as an intermediate step to determine cell memberships in the cluster level graph abstraction formulated by VIA. VIA offers a uniquely high degree of flexibility in terms of transitions between cell states making it possible to capture non-linear and non-tree behaviors and to avoid oversimplifying assumptions that bias the prediction of possible cellular transitions.

Chapter 4 showcases how image based features from imaging cytometry provide a new and promising basis to analyze cellular heterogeneity. While scRNA-seq is generally accepted as a gold standard for biological discovery and validation, the purpose of this chapter is to show that biophysical and morphological features of cells extracted from single-cell images can also be used as valid inputs to downstream analysis tools to delineate cell types (clustering and classification) as well as predict the sequence of progression in differentiation (unsupervised trajectory inference). As a proof of concept we demonstrate the application of unsupervised learning in terms of clustering (PARC) and trajectory inference (VIA) on imaging datasets where the “true” labels of cell type and stage are unknown to the algorithm. We also show an example of a supervised learning model developed for the classification of lung cancer cells where transfer learning improves the accuracy of classification across batches.

Chapter 5 discusses the scope for future work based on the published materials presented here.

Chapter 1: Introduction to single cell omics

Technological advances allow us to isolate individual cells and profile mRNA, chromatin accessibility, proteins and physical morphological traits at the single-cell resolution. The characteristics of these measurements stand in sharp contrast to bulk measurements of the previous decade that were examined using analyses that aggregated or averaged the data, and inevitably concealed or only partially characterized certain biological phenomena [Lee et al., 2020, Richa et al., 2021]. High-throughput experiments which profile transcriptomes, proteomes and cell morphology at the single-cell level necessitate advances in computational methods to unlock the knowledge contained in these datasets. A new computational paradigm has emerged in which algorithms must handle cells as stochastic units, capable of inter-cellular variations even within sub-cell types and therefore best described probabilistically. New methods need to navigate higher levels of feature dimensionality and simultaneously interrogate thousands of genes, or hundreds of proteins or biophysical features in this high-dimensional space. These present new challenges specific to the field of single-cell omics.

A single experiment generates a large volume of high-dimensional raw data. For instance, in the case of scRNA-seq, the stochastic expression of (tens of thousands of) genes for each of the thousands of individual cells is measured. In the case of imaging and flow cytometry data the sample size easily scales to 100,000s-millions whilst the data dimensionality of 10-100 features or surface markers remains comparatively lower than scRNA-seq data. Although we are seeing the emergence of scRNA-seq data approaching sample sizes of 100,000s-million, it is still more common to deal with a scenario where the number of genes exceeds the number of cells. This poses a unique set of problems related to selecting and filtering genes, handling noise and zero-inflated data and the curse of dimensionality. Conversely, in cytometry data, there may be an insufficient number of features to capture the true heterogeneity in the cell population with distinct populations merged together due to a lack of adequately discriminatory markers or morphological features. A range of quality control, normalization and dimensionality reduction methods are used to render these “big data” cell samples more manageable and comparably packaged for subsequent complex downstream analysis methods applied to distill information and carry out a variety of tasks: accurately identify rare or novel cell types, define distinct lineages and characterize diverse processes in development and pathogenesis [Richa 2021], all while overcoming the technical noise associated with millions of single stochastic measurements.

All single-cell analysis frameworks share some building blocks in common [Luecken et al., 2019] (see Figure 1.1). In order to interrogate the data, one has to first conduct some basic pre-processing steps such as quality control, normalization and dimensionality reduction. After this comes any combination of relevant downstream analyses for visualization, clustering, trajectory inference, differential analysis and compositional analysis.

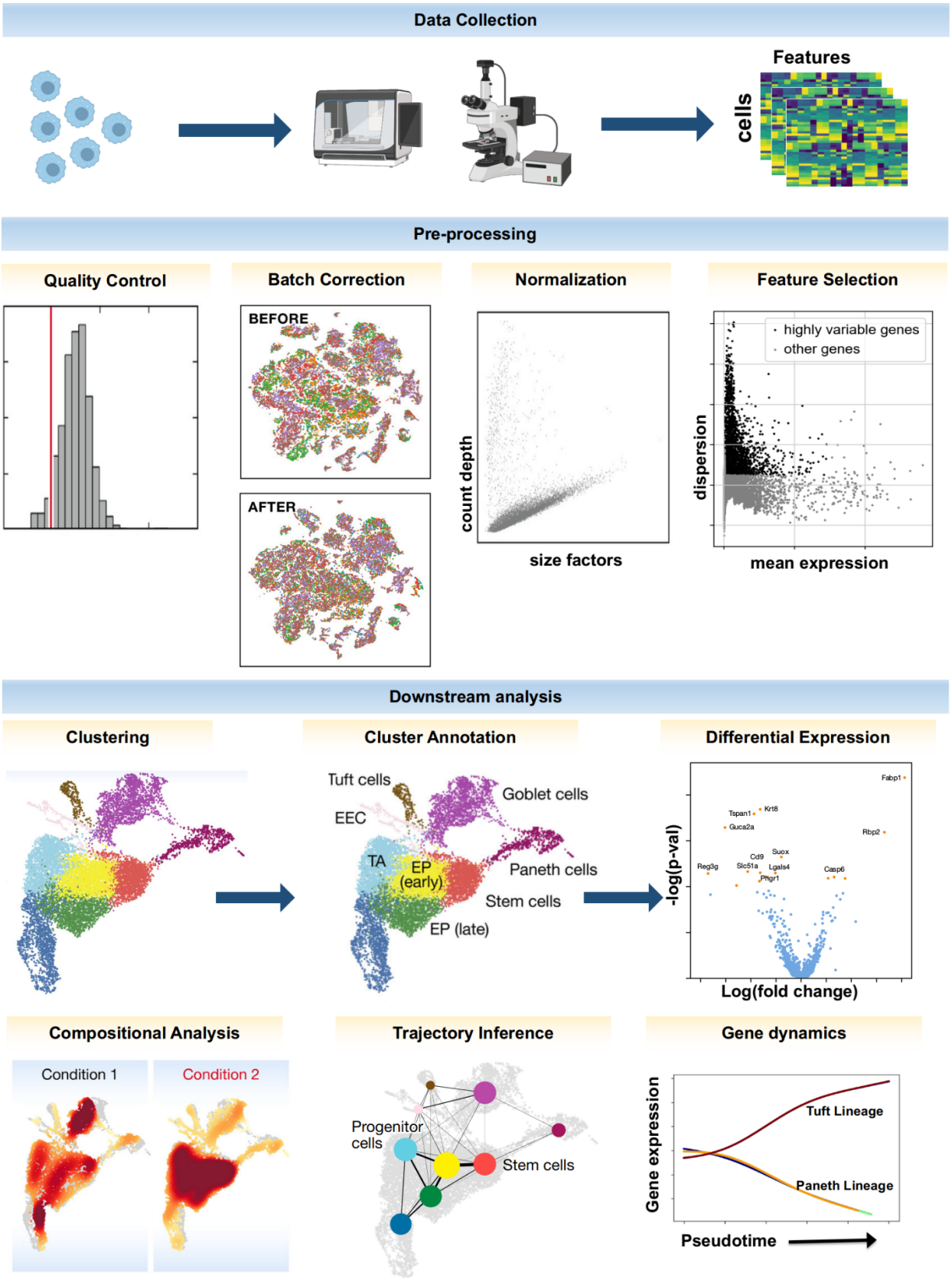


Figure 1.1 Single cell analysis pipeline adapted with modifications from Luecken & Theis 2019

Basic quality control filtering removes damaged or doublet cells. In the context of scRNA-seq this can mean filtering out cells that show very low or high count-depth as this can be indicative of quiescent cells or doublets, respectively. Similarly cells with a high fraction of mitochondrial gene expression might indicate damaged cells. In imaging and flow cytometry, gating strategies based on ratio of height and width of the forward and side scatter are commonly used to remove doublets. Protein dyes are used to viably remove (dying) cells with compromised membranes. The thresholds set for QC may vary between experimental setups.

There are several different normalization strategies, with the common intention of ensuring that downstream comparisons of feature expression between cells are valid. Simple Z-score normalization of features prevents any given feature from skewing the analysis based on magnitude and is often used in (mass, flow, imaging) cytometry data. There is currently no consensus on whether or not to perform normalization over genes in scRNA-seq data [Luecken et al., 2019] and researchers will typically try a few approaches to see which works best for their data. For scRNA-seq however, normalization of the scRNA-seq counts is a critical step that corrects for cell-to-cell differences in capture efficiency and sequencing depth. The simplest approach is library normalization which involves scaling the counts of each cell to remove any cell-cell differences in terms of total count. Transformations like log(1+x) may be optionally applied to reduce the skewness of the data and approximate the assumption of many downstream analysis protocols that the data is normally distributed.

Dimensionality reduction is often performed after data correction and prior to downstream analyses. It is carried out for two different objectives: the first is summarization and the second is visualization [Luecken 2019]. Visualization tries to optimally describe the data in 2-3 dimensions so that it can effectively be conveyed on a scatter plot. Summarization can retain a much higher number of dimensions that are usually ranked in terms of capturing variance in the data. In mass cytometry data where the number of dimensions is limited to the 20-50 surface markers being probed, dimensionality reduction for the purposes of summarization may not even be necessary, though it might be useful to summarize information if a subset of markers is highly correlated. ScRNA-seq data on the other hand, with tens of thousands of genes, suffers from the curse of dimensionality where all cells begin to look equidistant in space at very high dimensions and therefore cannot be effectively separated using distance metrics (which form the basis of most clustering and graph building protocols). The computational cost of handling 1000s of features without any dimension reduction is also overwhelming. In practice, not all genes are required for meaningful classification of cellular expression profiles, and dimension reduction enables us to capture the relevant biological signals in a more concise set of features. Perhaps the most commonly used technique for summarization is Principal Component Analysis (PCA). PCA is a linear projection method that linearly transforms the original dataset into PCs ranked in decreasing order of variance. It is computationally efficient and removes redundant features. While the first few components of a summarization method can be used for visualization, one usually gets much better representation by using a method dedicated to visualization. The converse is also true, and it is not advisable to use visualization components as the input to downstream analyses such as Trajectory Inference (TI) or clustering. Two widely used popular visualization techniques are t-Distributed Stochastic Neighbor Embedding (t-SNE) Uniform Manifold Approximation and Projection (UMAP) which are both graph-based nonlinear techniques that can be applied to a set of PCs to retrieve a more intuitive 2 (or 3) dimensional visual representation of the data. The main issue is that they tend to distort or lose global information and should

therefore be interpreted with a grain of salt. They are also far more computationally expensive than PCA, which is typically a preliminary step before using UMAP/t-SNE.

After running through the aforementioned pre-processing steps, the data is now suitably packaged and ready for the downstream analyses that yield biological insights. Whether we start with scRNA-seq or cytometry data, the input provided to downstream analysis methods after various pre-processing steps will comprise a large number of cells, spanning ~50-100 features or dimensionality reduced components. In the case of scRNA-seq data, these features are a dimensionality reduced reperesnetation of the 1000s of genes initially measured, and are supposed to capture the variance, connectivity or discriminatory information of the original data in a concise form.

We only provide a brief explanation of some different categories of downstream analysis as chapters 2 and 3 are dedicated to clustering and trajectory inference respectively. Clustering is a classical unsupervised task to categorize cells into groups based on similarity in expression profiles which are computed based on the distance metrics. The results can be of significance on their own to annotate cell types and markers, or can serve as a covariate in other downstream analyses. For continuous differentiation processes, however, discretization may not adequately capture cellular relationships or transitioning populations. Instead trajectory inference will try to reconstruct pathways such that cells progress along a trajectory that minimizes expression changes along each step of the way. This allows us to predict the temporal dynamics of lineage specific genes. Compositional and differential analyses probe the differences in population composition between two experimental settings, or the identification of genes that are significantly more or less expressed between clusters/settings. Areas such as data integration (across batches, experimental modalities) are also emerging to meet the needs of cross-modality experiments which are on the rise and prompt the need to be able to effectively aggregate experimental data and contribute towards atlas initiatives.

In the following chapters we examine in greater detail the current challenges related to Clustering and TI approaches and new approaches, namely PARC and VIA, which attempt to resolve some of these issues.

Chapter 2: Clustering

2.1 Introduction to clustering

In the context of single cell bioinformatics, we can consider the cell to be a fundamental unit of computation. Unsupervised clustering plays a decisive role in facilitating downstream biological interpretation of these fundamental units and can be described as the natural grouping of cells based on the similarity of attributes (physical such as cell size and shape, or surface markers or the transcriptomics) [Kiselevet al., 2019]. The term 'unsupervised' refers to the absence of a ground truth label. The discretization of cell types not only aids in cell type annotation and cell type discovery, but can also be a useful intermediate step to other downstream analyses such as trajectory inference and compositional analysis. In Chapter 3 we take advantage of the ability to synthesize cluster level data together with single-cell connectivity information to delineate continuous differentiation processes between sub-cell types.

The creation of cell atlas projects which aim to categorically map all cell types at different stages of development across organs [Jia et al., 2018, Bastidas-Ponce et al., 2019], embryos [Pijuan-Sala et al., 2019, Cao et al., 2019] and entire organisms [Tabula Muris Consortium 2020, Regev et al., 2017] has been spurred by the rapid proliferation of large scale single cell datasets. These atlases will advance our understanding of basic biology in terms of providing a reliable reference library and ontology for categorizing emerging data and the discovery of new cell types, and for studying diseases. The availability of computationally efficient and data driven clustering methods will be a necessary prerequisite for ensuring practical usability of these atlases and mapping clusters (cell types) from new datasets onto established cell types in the reference atlases [Kiselev et al, 2019].

A pressing issue in most existing clustering methods is the lack of a scalable data-driven capability to parse large and heterogeneous data. Most tools, especially those developed for gene expression data, become computationally prohibitive when the cell count exceeds 10^5-10^6 cells, which in turn also deters efficient parameter tuning required for hypothesis testing. Below, we provide an overview of some of the main clustering methods, their shortcomings, and new efforts to alleviate some of these challenges in order to harness the information potential in large scale single cell omic data.

2.2 Clustering methods and shortcomings

2.2.1 Common clustering methods

Most clustering methods fall into a few different categories. The most common is perhaps *k*-means where each cell is assigned to one of *k* iteratively defined centroids (cluster centers) and has the advantage of being very fast as it scales linearly with the number of cells.. However, major drawbacks of *k*-means are that it i) requires the user to predetermine the number of clusters to be identified and ii) demonstrates a tendency to derive equally sized, spherical clusters which tend to conceal rarer cell types. Methods like RaceID and SC3 use modified or extended versions of *k*-means to alleviate the issue of hiding rare populations. However, the overhead of ensuring small populations are not obscured by larger ones and

trying to identify the optimal k value in a data driven way increases the runtime to the extent that SC3 and RaceID require 5-6 hours to parse a scRNA-seq dataset of only 6000 cells.

Finite mixture models (e.g. the Gaussian mixture model) can help to overcome the spherical bias of k-means by taking into account variance along different dimensions and therefore capture convex clusters. However, whilst this is a relaxation of k-means's spherical clusters, mixture models do not effectively capture irregular or asymmetric clusters. FlowPeaks [Y.Ge et al., 2012] is a notable example of a clustering method that combines k-means and mixture models. However, the runtime for computing the number of components and their parameters is very long and often results in clustering that obscures several distinct cell types due to internal metrics favouring coarser clustering. Hierarchical clustering is also popular and intuitive, taking the form of either divisive (top-down) or agglomerative (bottom-up) clustering. It however scales with $O(n^2)$ and is therefore impractical from a runtime or memory perspective for large datasets.

Graph based community detection models are a promising avenue where quality measures such as modularity [see Eq. 2.1 below] of cell type groupings are maximized for a network constructed based on an exact or Approximate Nearest Neighbor (ANN) graph of the cells. They offer more flexibility in terms of cluster shape and size, and have the potential to be scalable. Since community detection algorithms form part of the PARC pipeline, we will provide a more detailed background of how this class of methods works and point out areas that are current weak points or bottlenecks.

2.2.2 Community detection algorithms

Modularity optimization of a network is an NP-hard problem, and many heuristic algorithms have been developed to accelerate the process. The Louvain method was until very recently by far the fastest algorithm [Blondel 2008], with the Leiden algorithm [Traag et al., 2019] recently emerging as a competitor. The combination of graph construction followed by Louvain community detection was first applied to single-cell (RNA-seq) data by Phenograph [Levine et al., 2015] and has also been incorporated into Seurat3 and Scanpy libraries. Whilst graph based approaches certainly have the potential to address many of the previously mentioned challenges, the existing offering still experiences a runtime bottleneck when handling datasets in the order of 10^5 and 10^6 cells. The unsatisfactory runtimes have perpetuated the issue of inefficient and slow parameter selection (even though parameter tuning in this case is more data driven, influenced by various parameters rather than directly predefining k number of clusters). Furthermore, rare or smaller cell sub-type populations are still often merged or subsumed by larger cell types for two interconnected reasons: i) the kNN graph construction phase forces a minimum number of neighbors onto each cell and can thus create false connections, ii) the resolution limit problem restricts the granularity of clusters that can be detected and becomes more pronounced as the number of cells increases - impacting large datasets which are of particular interest to us.

An explanation of the Louvain and Leiden methods is required to elucidate why spurious links can easily result in undesirable merging of clusters. The Louvain method optimizes modularity, which can be intuitively understood as assigning cluster membership such that the difference between the number of edges within a cluster and the expected number of edges in a random graph is greatest. For weighted networks, the modularity is defined as [Blondel et al., 2008]:

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta\left(c_i, c_j\right) \qquad \text{[Eq. 2.1]}$$

In Equation 2.1, A_{ij} is the weight of the edge between vertex i and j, k_i is the sum of vertex i's edge-weights. c_i is the community to which vertex i is assigned. m is the weight of the graph. The Kronecker delta function is 1 if c_i=c_j, and otherwise 0.

However, modularity optimization often exhibits resolution limit issues related to the unwanted merging of clusters due to spurious edges (further explained in Section 2.3.5). This is particularly notable in large scale networks for certain types of quality functions such as the commonly used one shown above (which is the default in both Leiden and Louvain implementations). It was recently noted that in addition to the resolution limit issue, the Louvain method can sometimes result in clusters with poor intra-cluster connectivity. A cluster might actually consist of two very weakly connected subclusters that should have been separated, but due to the greedy (locally optimal) nature of the algorithm, were not. After each iteration in Louvain, the nodes belonging to a cluster are merged such that they cannot be isolated again in subsequent rounds.

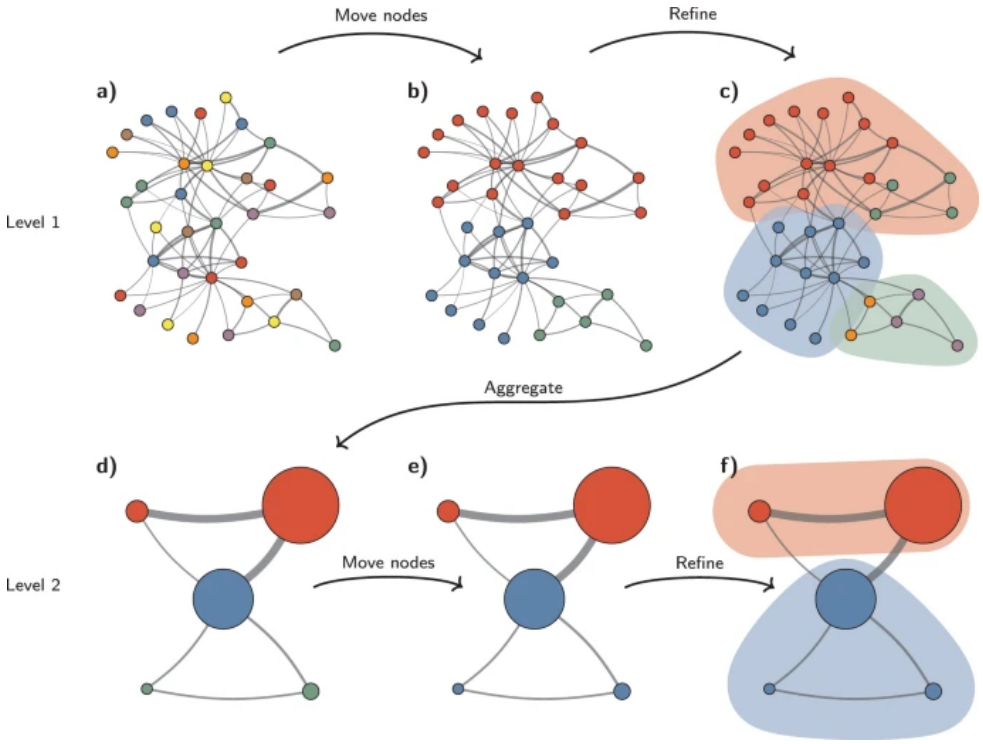


Figure 2.1 Key steps in the Leiden algorithm (a). The algorithm moves individual nodes from one community to another to find a partition (b), which is then refined (c). An aggregate network (d) is created based on the refined partition, using the non-refined partition to create an initial partition for the aggregate network. For example, the red community in (b) is refined into two subcommunities in (c), which after aggregation become two separate nodes in (d), both belonging to the same community. The algorithm then moves individual nodes in the aggregate network (e). In this case, refinement does not change the partition (f). These steps are repeated until no further improvements can be made. [Adapted directly from Traag 2019].

Leiden addresses the issue of internally disconnected communities by treating loosely connected parts of clusters as separate sub entities even though they are assigned membership to a common parent cluster

found in the aggregation step immediately before. This is intended to prevent the weak connectivity within clusters whilst still controlling the proliferation of clusters. However, in terms of the actual clustering membership solution, it means that once a sub entity forms, it can only be reassigned to one of the existing parent communities and does not exist as its own community.

The susceptibility to a resolution limit and poor intra-cluster connectivity which consequently diminishes the discriminative power of cluster labels, make the clustering solution susceptible to the presence of spurious edges in the graph representation of the cell-expression matrix resulting from a graph construction step that forces a floor (K nearest neighbors) on the number of edges extending from a cell. This will be one of the key issues we try to address in PARC and described in subsequent sections.

2.2.3 Challenges for efficient parameter tuning of clustering algorithms

Clustering algorithms offer different parameters to tune the final number of clusters presented in the clustering solution. *K*-means and variants require that the user presets the number of clusters (less data driven), whilst community detection methods use a parameter like the number of Nearest Neighbors to indirectly influence the number of clusters presented by altering the number of edges (and hence connectivity) introduced in the graph. One avenue to achieve a data-driven approach (which reduces user-bias) is to rely on internal clustering evaluation metrics (which do not require a ground truth reference) such as the silhouette coefficient to quantify a cluster quality score that can aid in the determination of how many clusters the algorithm should capture. These metrics, however, tend to favour a fairly coarse clustering that may overlook rare cell types or distinct sub cell types [Kiselev et al., 2019, Duo et al., 2018]. Practically speaking, the computation of said quality scores is also not computationally feasible for large data sets.

It is therefore preferable for clustering methods to be fast and efficient enough to allow researchers to test a range of parameters as different levels of granularity are required for different tasks (e.g. rare cell detection or delineation of groups to reduce the complexity of data prior to other downstream analyses). Intuitive parameters will also allow researchers to use their judgment to identify meaningful clusters that consistently emerge across a reasonable range of parameters. The stability of certain cell groups which persist across parameters may serve as a form of hypothesis validation that the categorization corresponds to biologically distinct cell types and not algorithmic artifacts.

2.3 PARC

2.3.1 Introduction to PARC

In view of these challenges, particularly with respect to runtime on larger data, the difficulty of capturing irregular cluster shapes or segregating smaller populations, and the difficulty of probing the desired number of clusters in a data driven, intuitive and yet computationally efficient matter, we developed a new graph-based clustering pipeline PARC, Phenotyping by Accelerated Refined Community-partitioning. PARC is a fast, automated, combinatorial graph-based clustering approach that integrates hierarchical graph construction and data-driven graph-pruning with a community detection algorithm. PARC's hierarchical graph-pruning step enables it to (i) outperform existing tools in scalability (>1 million cells with wide range of dimensionality) and (ii) augments the sensitivity and specificity to unbiasedly reveal the cellular heterogeneity, especially rare subsets within large populations. We validate the performance of PARC on large-scale datasets, with respect to speed and accuracy, as well as versatility across a wide range of single-cell data including: mass and flow cytometry, scRNA-seq and imaging cytometry (Fig. 2.5 -Fig 2.7 as well as extensive benchmarking presented in Stassen et al., 2019). Notably, we demonstrate that PARC can detect subpopulations that were not labeled in the original scRNA-seq datasets of 68 000 peripheral blood mononuclear cells (PBMCs) [Zheng et al., 2017]. PARC also enables fast data driven clustering of the mouse brain dataset of 1.3 million cells.

2.3.2 Overview of PARC methodology

We introduce three key steps employed by PARC to enable scalable and data-driven clustering of single-cell data (Fig 2.2).

1. The first step is an accelerated nearest-neighbor graph construction using hierarchical navigable small world (HNSW) [Malkov & Yashunin, 2016], in which each node is a single cell connected to a neighborhood of its similar cells by a group of edges.
2. The second step is the data-driven pruning of the edges based on the statistical distribution of edge-weights at both the local node-by-node level and the global network level to remove spurious linkages (which deter cell type isolation) and redundant edges (which slow down the community detection). This is a key step in speeding up the community detection and enabling the detection of minor populations.
3. The last step is community detection based on the Leiden algorithm [Traag et al., 2019] with modifications made to efficiently handle singletons (clusters containing one or very few data points) resulting from the pruning and to optionally further cluster selected communities. The modifications in terms of singleton handling prevent PARC from encountering errors across a wider range of parameters. These steps are integrated in such a way that PARC's performance is not determined by each individual step, but the feedback between them.

Notably, the pruning procedure in PARC, which reduces the sample size of edges and improves the fidelity of the KNN graph representation to the underlying data, critically increases the speed and robustness of the subsequent community-detection step. We find that this is particularly advantageous in detecting rare but distinct populations as it shields against the resolution limit problem. In Sections 2.3.3-2.3.5, we will describe in detail the three modules and their integration.

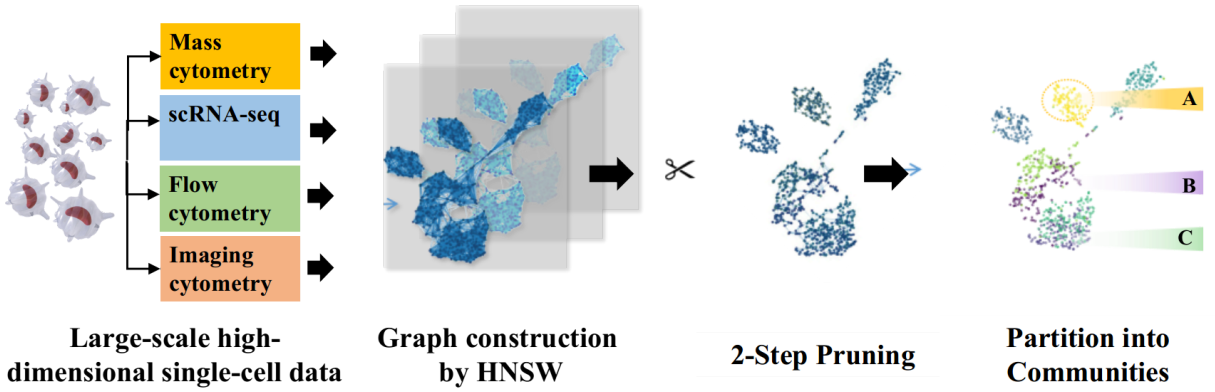


Figure 2.2 Overview of PARC workflow. PARC can be used for large-scale single-cell analysis on multiple types of high-dimensional single-cell data. The enabling features include fast graph construction by HNSW, 2-step data-driven graph refinement and pruning, and accelerated community detection by Leiden algorithm. Adapted from Stassen 2020.

2.3.3 Graph Construction using HNSW for fast and scalable KNN search

In the first step, PARC receives as an input the cell-feature matrix (features can be signal intensity, cell-level measurements or Principle Components of the gene expression count matrix) and constructs a KNN graph using an Approximate Nearest Neighbor (ANN) algorithm for a Hierarchical Navigable Small World (HNSW) KNN graph [Malkov and Yashunin, 2016]. Approximate Nearest Neighbor algorithms can take a fraction of the time required by exact neighbor searches whilst only marginally sacrificing accuracy. There are several competitive ANN methods so our selection of a method factored in maturity of code, availability of API across different OS, being lightweight in terms of dependencies, and speed and recall in both the construction and querying phase. Furthermore, the fundamental structure of small world graphs is well suited to our use case since we expect our data to comprise distinct clusters that may be loosely interconnected and thus mirror the intended structure of HNSW. A small world graph is characterized by long links which bridge different clusters allowing non-neighbor nodes to be accessed by traversing relatively few edges, and shorter links between nodes whose neighbors are likely to be neighbors of each other and thus represent intercluster connectivity. The HNSW method differs from other navigable small world methods by binning links in hierarchy (i.e. layers) according to their lengths. The search starts at the top layer containing the longest links and traverses the elements until a local minimum is reached. The search then goes to the lower layer (i.e. the layer having shorter links) from the node where the most recent local minimum was detected. Such hierarchical graph structure allows fast graph construction and query.

2.3.4 Pruning edges to ensure effective capture of network structure prior to clustering

At this stage we could feed the HNSW based KNN graph to a community detection method and see some reasonable gains in scalability compared to other methods (Fig 2.4). However we would like to explore the possibility of further improving the speed and accuracy of the clustering by pruning the graph edges. In PARC, a hierarchical pruning method is applied before clustering that is critical for both the speed at which communities can subsequently be established and for the identification of subtle cell types without incurring fragmentation.

As mentioned earlier, the clustering solution is a direct result of the quality of edges in the network and care must therefore be taken to ensure that the edges are curated to be as faithful to the underlying biology as possible. One common strategy to impact the graph connectivity and improve the clustering is to tune the user-defined K value (K number of nearest neighbors). Higher K values generally favor preserving larger communities, but compromise the ability to detect rare subpopulations. On the other hand, lower K values in other clustering methods are only marginally (and inconsistently) better at recovering rare populations but can cause over-fragmentation—complicating the biological discovery.

In PARC, we pursue a pruning strategy motivated by the observation that the edge-weight statistics in various single-cell datasets commonly exhibit a long-tailed distribution.

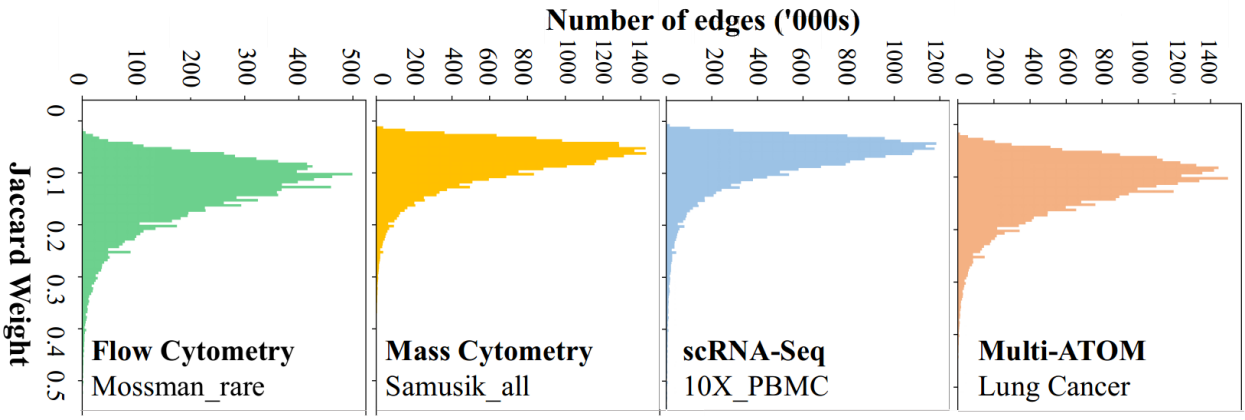


Figure 2.3 Distributions of graph edge-weights in various single cell datasets. The high weight score of important neighbors in the tail diminishes the difference between weak and majority links negatively impacting the robustness and speed of community detection – an issue that could be addressed by graph pruning. Adapted from Stassen 2020.

The skewness in the distribution means that the relative weight difference based on Jaccard similarity (and also Euclidean distance) between the weak and majority edges is diminished due to the fact that the long tail occupies a large portion of the scale but a small portion of the actual cell population. However, this problem, conceivably a result of the 'curse of dimensionality', cannot be solved by simply re-weighting the graph using a different metric as it is a direct function of the dimensionality of the data. Consequently, the optimization function employed in the subsequent community-detection step sees the weak (potentially spurious) and majority edges as very similar in importance which confounds the community partitioning.

To address the limitations posed by edge-weighting and selecting a suitable number of nearest neighbors to initialize the graph, PARC instead starts with a generous fixed K number and implements automated two-step pruning of weak edges guided by the data structure. First, it examines each node locally and removes the weakest neighbors of that particular node based on the Euclidean distance; and second, it re-weights the edges using the Jaccard similarity coefficient and globally removes edges below the median Jaccard-based edge-weight. The local pruning allows us to remove redundant neighbors in very densely connected neighborhoods, whereas the global pruning removes spurious edges that would otherwise persist in more sparsely connected regions. We compare the runtime break-down between the graph-based algorithms PARC, Phenograph and Seurat in terms of network construction and modularity

optimization steps in their default settings. As shown in Figure 2.4, the impact of pruning becomes more pronounced in lowering clustering runtime when sample size increases.

The aggressive pruning in PARC can generate several small clusters or singletons which are not necessarily all outliers and need to be returned to the appropriate parent community based on a consensus vote of single-cell neighborhood relationships. PARCs efficient handling of fragments overcomes prohibitive runtime bottlenecks such as those experienced by Phenograph and Seurat (when lowering K). Sensitivity analyses showed that the pruning not only increases the accuracy of rare cell detection without incurring fragmentation in other clusters, but that it also extends the range of user chosen values of number of nearest neighbors in which good performance is achieved. Generally speaking, due to its fast runtime, users can efficiently configure parameters in PARC if they wish.

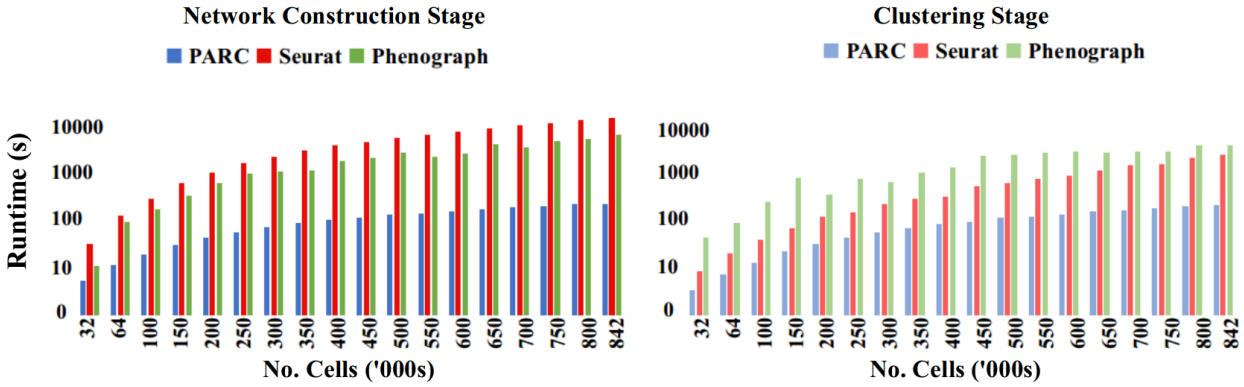


Figure 2.4 PARC runtime and scalability tests: Runtime comparisons of PARC, Phenograph and Seurat in terms of graph construction and clustering time on random samples of CyTOF data. Adapted from Stassen 2020.

2.3.5 Pruned graph helps shield against resolution limit and undesirable mergers

Having constructed a network representation of the cells, we apply the Leiden modularity optimization algorithm (Traag et al., 2019). The pruning step alleviates issues around the resolution limit. To explain why this is the case, we need to examine the modularity quality function of the graph—a measure of density of links within a community to that between communities. Only moves that yield a positive change in modularity are carried out. If we assign all nodes in community A to B, then the change in modularity is:

$$\Delta Q_{AB} = \sum_{i \text{ in } A} \frac{k_{i,in}}{m} - \frac{k_i k_{B,tot}}{2m^2} \qquad \text{[Eq 2.2]}$$

Here $k_{i,in}$ is the sum of weighted links from node i to nodes in community B, k_i is the weighted links incident on node i, $k_{B,TOT}$ is the sum of weighted links incident on B, m is graph weight.

For the simplified case of an unweighted graph (or a graph where the weightings are not discriminatory and hence effectively unweighted), we rewrite the change in modularity when merging community A and B as (where k_A and k_B are the total degrees of A and B, and L is the total number of links in the entire network, and l_{AB} is the number of links between community A and B; Barbasi, 2019):

$$\Delta Q_{AB} = \frac{l_{AB}}{L} - \frac{k_A k_{B,tot}}{2L^2} \quad \text{[Eq 2.3]}$$

Consider the scenario where $k_A k_B/2L < 1$, then the change in modularity is positive if there exists even one link between the two communities $(l_{AB} >= 1)$. For the sake of simplicity, $k' = k_A \sim k_B$, then $\Delta Q > 0$ when A and B are merged for all $k' \leq \sqrt{(2L)}$. Therefore, if the number of links within a small community is below the threshold $\sqrt{(2L)}$, then a link to another community will incur a merger and the algorithm will struggle to resolve communities below the resolution limit of $k' \leq \sqrt{(2L)}$. It is therefore critical to remove artificial or weak links set up in the initial KNN graph.

2.4 Results

2.4.1 PARC's pruning step enables identification of rare populations in cytometry data

The impact of pruning, which is a key algorithmic contribution in PARC, is tested with respect to isolating rare populations in three cytometry datasets of 40,000-400,000 cells each, out of which the rare cell type represents less than 0.1% of the total population. The rare populations in the flow cytometry datasets (Nilsson and Mosmann) are annotated based on manual gating and have been used previously in benchmarking studies [Weber & Robinson 2016] whilst the individual cell type samples in the image-based Multi-ATOM dataset were digitally mixed. We consider the performance of the three graph based methods PARC, Phenograph and Seurat when resorting to lowering the K parameter (number of nearest neighbors) as a potential alternative solution to hierarchical pruning in order to segregate rare populations, the cluster with the highest F1-score for any cluster containing members of the rare population is reported. However, as shown in the heatmap Figure 2.5 this is an ineffective remedy for PARC, Phenograph and Seurat, and also leads to over-fragmentation of clusters that confounds downstream analysis. The better strategy is to enable PARC's pruning step (Left most column K=30, "Prune") which successfully segregates the rare cells.

A detailed sensitivity analysis of tuning the pruning parameter in PARC is provided in the Supplementary Materials (Stassen et al., 2019) and highlights ranges of reasonable values for pruning (given by the number of standard deviations from the mean edge weights) for several rare and multi-population datasets in terms of the F1-score, runtime and number of clusters generated. Generally speaking, Fig. 2.7 shows that pruning is more consequential for rare populations, however minimizing the level of pruning does cause a gradual decline in the ability to detect all subtypes within a multi-population dataset when the population size of these subtypes dwindles towards those of rarer populations. Given PARC's fast runtime, the parameter can be efficiently tuned, and the fragmentation handling mechanism in PARC controls the fragmentation of clusters which may otherwise impede further downstream analysis.

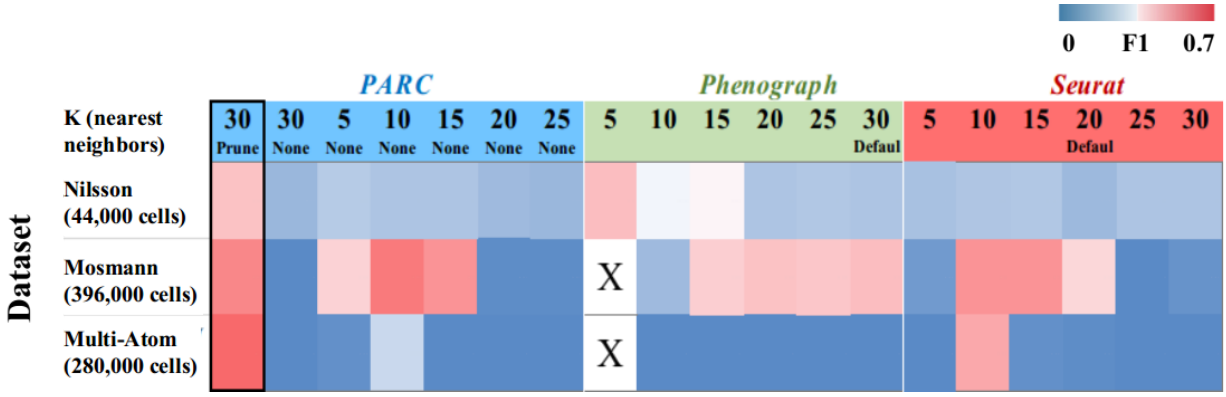


Fig. 2.5 Lowering the K value (number of Nearest Neighbors in KNN graph construction) to isolate rare cells is an ineffective strategy compared to graph pruning. Comparison of methods when lowering the value of K to show how this is an ineffective strategy for rare cell detection and to highlight that the pruning step of PARC (left most column K=30 "Prune") is more effective and works even when K is reasonably high (K=30, a common default level). The top header row indicates the number of K nearest neighbors and in the case of PARC shows the benefit of pruning ("None", means disable pruning, whereas "prune" means the default pruning step is applied). The heatmap displays PARC, Phenograph and Seurat accuracy results for identifying the rare cell population in 3 datasets: Nilsson_rare, Mosmann_rare, multi-ATOM_rare, with rare populations of 0.08%, 0.03%, and 0.04%. Legend indicates F1-score of the cluster with the highest F1-score for any cluster containing rare cells. X's denote stalled process due to no efficient fragmentation handling for low K in other methods. [Adapted from Stassen 2020]

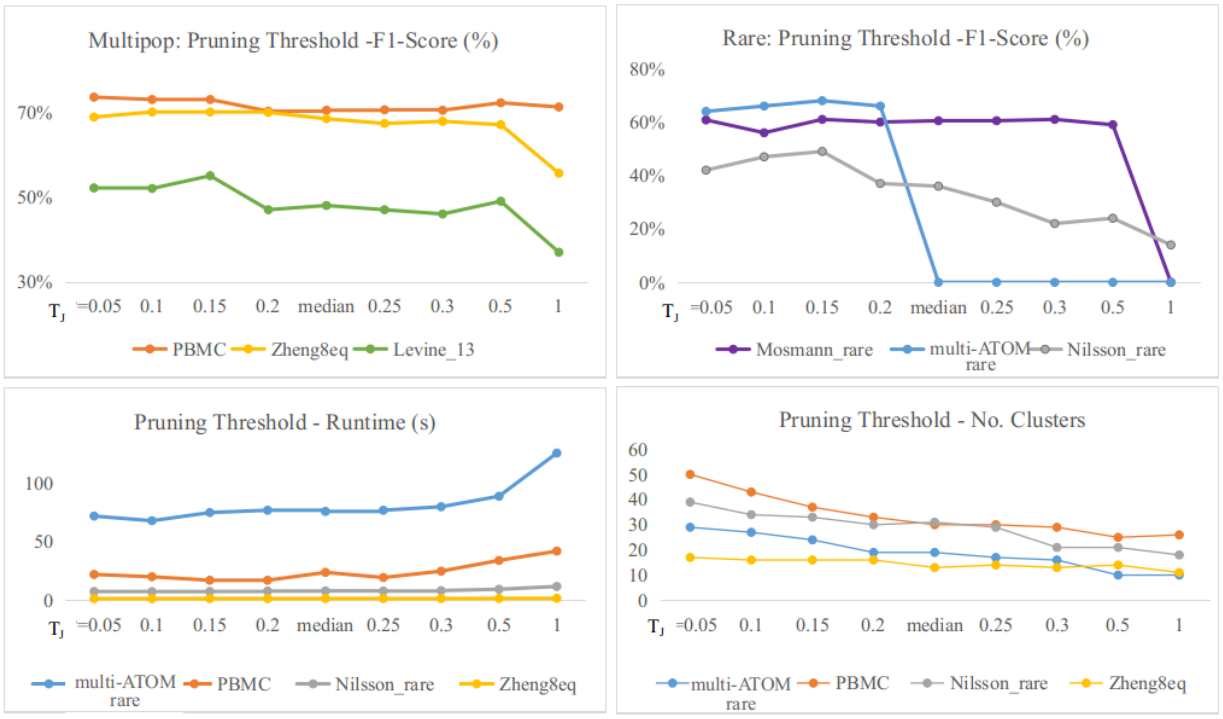
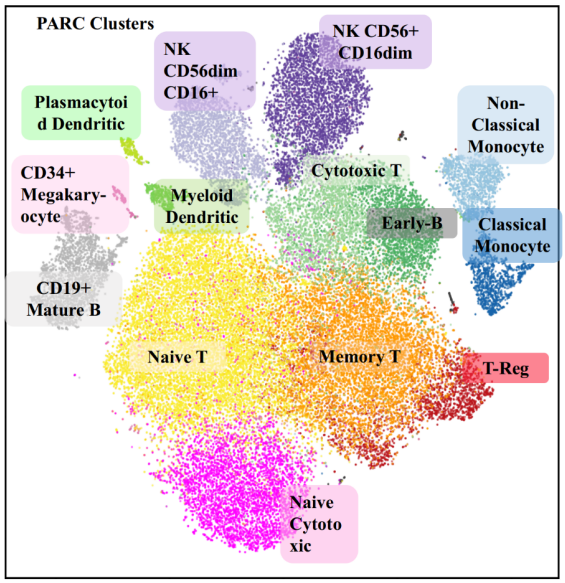


Fig. 2.6 Tuning the graph-pruning parameter in PARC. Accuracy on multi-population data and rare population data shows varying levels of dependency on pruning (top). The runtime is reduced by pruning, with efficient fragmentation handling ensuring that runtime bottlenecks do not emerge at high levels of pruning. The number of clusters generated remains reasonable for downstream analysis even at significant pruning levels (bottom). [Adapted from Stassen 2020 Supplementary Materials]

2.4.3 PARC used for large scRNA-seq profiling of cells

In addition to testing PARC on imaging flow and mass cytometry data (as shown in the examples of rare-cell detection from the previous section), we also tested PARC on various annotated scRNA-seq datasets and benchmarked their performance against popular methods highlighted in a study [Weber & Robinson 2016]. Accuracy and quality of clustering was quantified by the Adjusted Rand Index and the F1-score [Supplementary Fig. S9 in Stassen et al., 2019]. In this section we highlight the use of PARC on a mid-size scRNA-seq dataset of 68000 PBMCs to show an example of PARC in immune cell profiling. PARC identifies subpopulations that were masked by the original manual gating (Fig. 2.7a–c). This is attributed to the fact that the annotation was mainly given to T-cell subpopulations on a mesoscopic level (e.g. CD4+, CD8+, memory and regulatory T cells). In contrast, other subtypes of PBMCs (e.g. monocytes, dendritic cells and natural killer cells) are not annotated by any of their known subtypes. Nevertheless, PARC is able to reveal the clusters showing high expression of CD14 (cluster 9) and CD16 (or FCGR3A) (cluster 10), markers for classical and non-classical monocytes, respectively [Ong et al., 2018]. It also identifies subsets of NK cells as inferred by the expression level of CD160 and CD16 (FCGR3A) (clusters 3 and 5), which is known to be associated to the CD56dim CD16+ cytotoxic NK cell phenotype (cluster 5) [Le Bouteiller et al., 2011]. Notably, PARC also detects rare populations of IL-3RA+ [Zhang et al., 2017] plasmacytoid dendritic cells (cluster 11, 0.6%) and megakaryocytes (cluster 12, 0.4%).

(a)



(b)

Cluster	Inferred Cell type	Markers
11	Plasmacytoid Dendritic	IL-3RA, GZMB
13	Myeloid and Monocyte related CD14+ Dendritic	CD14, CD1C+, HLA-genes
9	Classical Monocyte	CD14+ FCGR3A- (CD16-), CX3CR1-, S100A12+
10	Non-Classical Monocyte	CD16+, CD14low, CX3CR1+
8	Mature B cell	CD22+, CD79A+, CD79B+, CD19+
17	Early-B cell	IGJ+, CD79A+
12	CD34+Megakaryocyte	PF4+,GP9+, TREML1+, PPBP+
6	CD25+ T-Reg	FOXP3+, CD52+
0,15,14	CD4+ Naive T	CD8A-, CCR7+
1	CD4+ T-Memory	CD8A-, CCR7-
16	CMV-specific CD4+ TMemory	IFI6+, IFI27+, IFIT3
3	NK II CD56brightCD16dim	CD160-, NKG7+
5	NK cytotoxic CD56dimCD16+	NKG7+,FCGR3A+(CD16+), CD160+,CD2-, CCR7-
2	CD8+ Naive Cytoxic	CCR7+ CD8A+
4,7	CD8+ CytoToxic (activated)	CD8A+, NKG7dim, GZMK

(c)

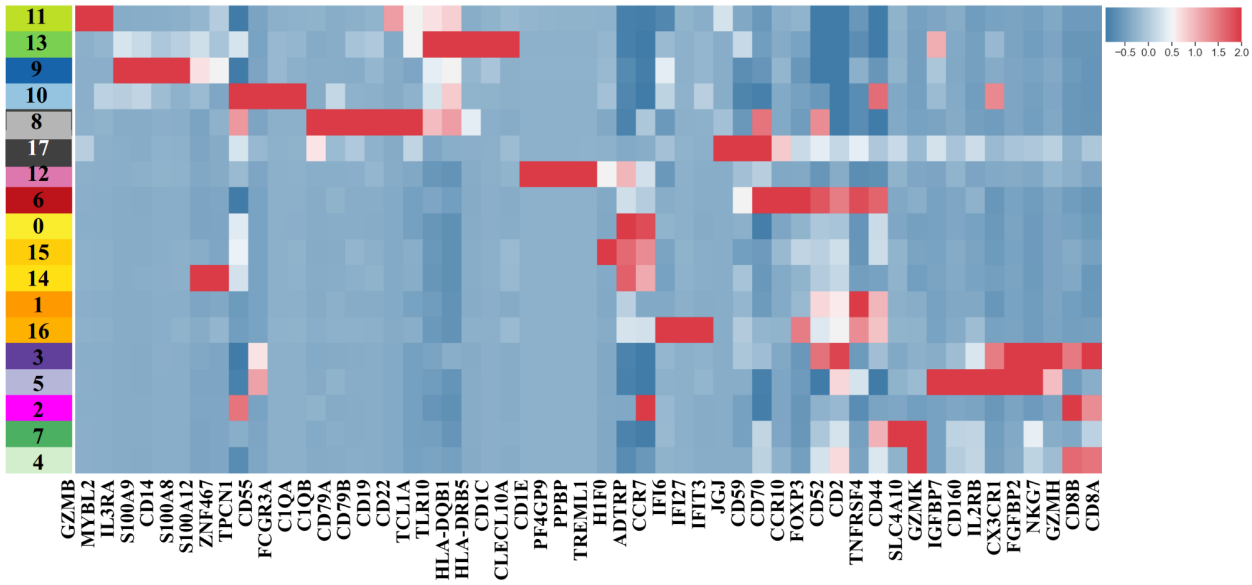


Figure 2.7 PARC for sc-RNA analysis of 68K PBMC (a) t-SNE visualization of 68K PBMCs (Zheng et al. 2017), colored based on PARC clusters, delineates well-known cell subtypes not captured in original annotation (b) table of marker genes (extracted from heatmap) used to infer cell type (c) Heatmap of most differentially (log2-fold) expressed genes in each cluster. Legend represents mean cluster-level z-score normalized gene expression.

2.5 Concluding remarks

In this section we highlighted key challenges faced by many clustering methods when analyzing high dimensional, large scale heterogeneous single cell data. We showed how PARC aims to address these gaps by employing a graph-based clustering approach that outperforms other methods not only in speed and scalability but also the ability to accurately capture data structure and detect rare populations. To deal with large-scale data processing, PARC does not incur prohibitive computational costs nor resort to data downsampling. Instead, PARC is built on three integrated elements: (i) HNSW for accelerated KNN graph construction, (ii) data-driven two-step graph pruning which enables the detection of rare cells and distinct cell subtypes and (iii) the community-detection Leiden algorithm augmented by PARC's KNNgraph based mechanism to handle small cluster fragments. Our results show that PARC's graph pruning step, guided by the local and global single-cell data structure, refines and improves the data graph representation which in turn accelerates Leiden community detection and alleviates the common problem of the resolution limit in community detection which otherwise impedes the delineation of smaller populations.

Since its publication, PARC has remained in the top 5% of research outputs rated by Altmetric and integrated in various academic and industry pipelines. The availability of tutorials to guide biologists has allowed easy adoption of PARC in many pipelines. Notably, PARC has been used to phenotype immune cell composition in PBMCs in various covid related studies. For instance PARC was used to study transcriptomic alterations in immune cell types in PBMCs of patients before and 4 weeks after inoculation, revealing a reduction in CD8+ T cells accompanied by an increase in classical monocytes [Liu et al., 2021]. Another interesting study that used PARC was for an extensive longitudinal mass cytometry profiling of CD8+ T cells in Sars-Cov-2 convalescent patients [Schulien et al., 2021].

Chapter 3: Trajectory Inference

3.1 Introduction to trajectory inference and current challenges

In chapter 2 we looked at automatically grouping similar cell types into discrete clusters. In this chapter we look at unsupervised approaches to predicting continuous differentiation processes. Understanding dynamical processes underlying cellular differentiation and lineage commitment to different cell types is a focus in stem cell and developmental biology. The ability to catalog changes in cell states and subsequently chronologically sequence the mechanism of choices made for cells in organs and tissues to move towards their fates during embryogenesis can provide cues on how to recapitulate this in vitro, and also shed light on the development of pathologies which arise as as a deviation from normal differentiation. Trajectory Inference methods are applied to the high-content readouts, such as those presented in single-cell omics data, to computationally predict the differentiation topology, and track the chronology of dynamical changes in molecular or biophysical signatures that give rise to the evolving cellular heterogeneity observed in tissue, tumor and cell populations.

These computational methods typically also calculate a "pseudotime" for every cell. This is a numeric value in arbitrary units which measures how far a particular cell is within a dynamic process of interest. By ordering the cells according to this pseudotime, it becomes possible to designate different transitional stages through which a cell progresses during its dynamic process. The initial dataset can be either a single snapshot of a mixture of cells in different stages or a set of samples collected at different timepoints. Starting from such a dataset describing high-dimensional, single-cell data, TI methods aim to order the cells with respect to an underlying dynamic process that explains the cell heterogeneity in the sample. Most TI methods will have one or two dimensionality reduction steps such as PCA (to first reduce the dimensionality from 1000s to 100s), sometimes followed by UMAP, t-SNE or diffusion maps (down to 2-3 components). At this stage most methods take one of two routes: The first being to cluster the single cells using K-means or hierarchical clustering and then connect these clusters using a Minimum Spanning Tree (MST) (which is of course quite restrictive in terms of topological configurations). The second approach is to skip any clustering and make a KNN graph of the single cells which increases the flexibility of transitions along the cellular landscape. However, oftentimes, in order to simplify the graph, a Minimum Spanning Tree is constructed at the single cell level. Once a single-cell KNN/MST or a cluster-level MST has been created, a shortest path algorithm (which is deterministic) is applied from a designated root cell to each cell in order to determine each cell's pseudotime in relation to the starting state. Further downstream analysis such as cell fate identification and lineage pathways typically requires manual intervention by the user and is another major drawback.

In this chapter we will introduce a new Trajectory Inference (TI) method VIA. VIA piggybacks on PARC's fast discretization of data as an initial part of its pipeline to predict the chronology of events occurring in differentiation processes. It uses a probabilistic graph based approach that tries to minimize the restrictions imposed on the graph structure and instead tries to retain complex transitions and topologies. VIA is an effort to overcome current challenges in TI. A recent benchmarking study of 45 TI methods [Saelens 2019] underlined the urgency for more accurate, scalable and user friendly approaches. The 45 methods typically follow the TI process described above and can be restrictive for various reasons.

Based on their evaluation we identified 4 challenges that need to be addressed to spearhead the next generation of TI:

1. First, it remains difficult to accurately reconstruct high-resolution cell trajectories and automatically detect the pertinent cell fates and lineages without relying on prior knowledge to adjust input parameter settings. Without automated cell fate and pathway prediction, downstream analysis tracking lineage specific traits and comparisons to other lineages in an unsupervised manner becomes difficult and is therefore a critical gap in many TI methods. However, even the few algorithms which automate cell fate detection (e.g., SlingShot [Street et al., 2018], Palantir [Setty et al., 2019], STREAM [Chen et al., 2019], and Monocle3 [Cao et al., 2019]) exhibit low sensitivity to cell fates and are highly susceptible to changes in algorithm parameters. Increasing the accuracy of cell fate prediction and lowering sensitivity to parameters would allow more unbiased hypothesis testing.
2. Second, current trajectory inference (TI) methods predominantly work well on tree-like trajectories (e.g., Slingshot and STREAM), but lack the generalizability to infer disconnected, cyclic or hybrid topologies without imposing restrictions on transitions and causality [Saelens et al., 2019]. This attribute is crucial in enabling unbiased discovery of complex trajectories which are commonly not well known a priori, especially given the increasing diversity of single-cell omic datasets.
3. Third, the growing scale of single-cell data, notably cell atlases of whole organisms [Tabula Muris Consortium 2020, Regev et al., 2017], embryos [Pijuan-Sala et al., 2019, Cao et al., 2019], and human organs [Jia et al., 2018, Bastidas-Ponce et al., 2019], exceeds the existing TI capacity, not just in runtime and memory, but in preserving both the fine-grain resolution of the embedded trajectories and the global connectivity among them. Very often, such global information is lost in current TI methods after extensive and multiple rounds of dimension reduction or subsampling that may be required by many TI methods to feasibly parse large datasets.
4. Fourth, fueling the advance in single-cell technologies is the ongoing pursuit to understand cellular heterogeneity from a broader perspective beyond transcriptomics. A notable example is the emergence of single-cell imaging technologies that now allow information-rich profiling of morphological and biophysical phenotypes of single cells, and thus offer mechanistic cues to cellular functions that cannot be solely inferred by proteomic or sequencing data. However, the applicability of TI to a broader spectrum of single-cell data has yet to be fully exploited.

3.2 VIA Method

3.2.1 VIA Introduction and algorithm overview.

To overcome these recurring challenges, we present VIA, a graph-based TI algorithm that uses a new strategy to compute pseudotime, and reconstruct cell lineages based on lazy teleporting random walks integrated with Markov chain Monte Carlo (MCMC) refinement (Fig. 3.1). VIA relaxes common constraints on traversing the graph, and thus allows capture of cellular trajectories not only in multi-furcations and trees, but also in disconnected and cyclic topologies. The lazy-teleporting MCMC characteristics also make VIA robust to a wide range of preprocessing and input algorithmic parameters, and allow VIA to consistently identify pertinent lineages that remain elusive or even lost in other top-performing and popular TI algorithms. 5 algorithms were shortlisted for comparative analysis conditional on meeting several of the following criteria: automated lineage path and cell fate prediction, recovery of complex topologies not limited to trees, scalability and generalizability to multiple single-cell-modalities. All methods were tested across a variety of transcriptomic, epigenomic, and integrated multi-omic datasets.

VIA's flexibility with respect to trajectory topologies and automated cell fate prediction allowed it, most notably, to detect two minor dendritic sub-populations and their characteristic gene expression trends in human hematopoiesis; identify pancreatic islets including rare delta cells; and recover endothelial and cardiomyocyte bifurcation in integrated data sets of single cell RNA-sequencing (scRNA-seq) and single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq).

Another defining attribute of VIA is its resilience in handling the wide disparity in single-cell data size, structure and dimensionality across modalities. Specifically, VIA is highly scalable with respect to number of cells (10^2 to $>10^6$ cells) and features, without requiring extensive dimensionality reduction or subsampling which compromise global information. Most TI methods require two stages of dimensionality reduction in the form of PCA followed by a subsequent stage of UMAP or MLLE down to just 2-3 components. While using 2-3 UMAP or t-SNE components for visualization can be an easy and intuitive way to present data, it is widely recognized to incur a significant loss of global information. Whilst any pre-processing is subject to researcher judgment, there is a general consensus that using PCA (or similar) to reduce 10,000s of sparse gene readouts to 100s of PCs is a sensible and information preserving step, the further reduction down to 2-3 dimensions may be too distortional and computationally unnecessary.

VIA Algorithm Overview

VIA applies a scalable probabilistic method to infer cell state dynamics and differentiation hierarchies by organizing cells into trajectories along a pseudotime axis in a nearest-neighbor graph which is the basis for subsequent random walks. Single cells are represented by graph nodes that are connected based on their feature similarity, e.g., gene expression, transcription factor accessibility motif, protein expression, or morphological features of cell images. A typical routine in VIA mainly consists of four steps: Accelerated and scalable cluster-graph construction, probabilistic pseudotime computation, automated lineage detection, visualization and downstream pseudo-temporal analysis of gene dynamics.

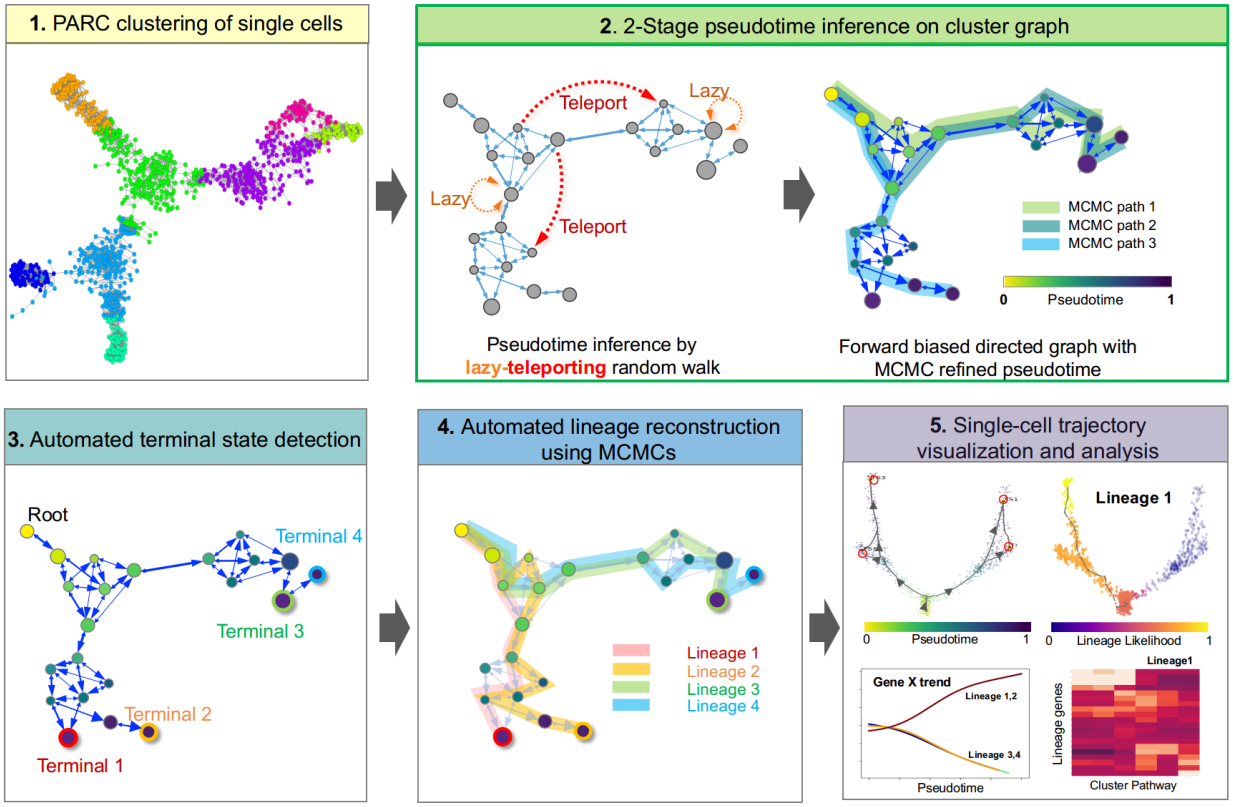


Figure 3.1 General workflow of VIA algorithm. Step 1: Representation of the single-cell data as a cluster-graph with each node as a cluster of single cells (computed by clustering algorithm PARC). Step 2: 2-stage pseudotime computation: (i) The pseudotime is first computed by the expected hitting time for a lazy- teleporting random walk along an undirected graph. At each step, the walk (with small probability) can remain (orange arrows) or teleport (red arrows) to any other state. (ii) Edges are then forward biased based on the expected hitting time (See the forward biased edges illustrated as the imbalance of double-arrowhead size). The pseudotime is further refined on the directed graph by running Markov chain Monte Carlo (MCMC) simulations (See the 3 highlighted paths starting at root). Step 3: Consensus vote on terminal states based on vertex connectivity properties of the directed graph. Step 4: lineage likelihoods computed as the visitation frequency under lazy-teleporting MCMC simulations. Step 5: visualization that combines network topology and single-cell level pseudotime/lineage probability properties onto an embedding using GAMs, and unsupervised downstream analysis (e.g. gene expression trend along pseudotime for each lineage).

3.2.2 Scalable cluster-graph construction and initialization of trajectory

VIA first represents the single cell data in a k-nearest neighbor (KNN) graph where each node is assigned cluster-level membership. The clusters are computed by our recently developed clustering algorithm, PARC [Stassen et al., 2019]. A cluster graph is then constructed where each node is a cluster of cells and the connectivity between clusters is determined using the initial single-cell level KNN graph. The cluster-level topology is an abstraction of the single-cell-level graph, and provides a coarser but clearer view of the key linkages and pathways of the underlying cell dynamics without imposing constraints on the graph edges. Together with the strength of PARC in clustering scalability and sensitivity, this step allows VIA to faithfully reveal complex topologies namely cyclic, disconnected and multifurcating trajectories (Fig. 3.2). It also smooths noise and speeds up probabilistic computation in subsequent steps of the algorithm. The Root node is defined in one of three ways: 1) automatically inferred if RNA-velocity is available 2) identified by the user or 3) jointly determined by user (as any node of a particular cell type) and fine-tuned by VIA based on node connectivity properties.

3.2.3 2-Step Probabilistic pseudotime computation

The trajectories are then modeled in VIA as: (i) lazy-teleporting random walk paths along which the pseudotime is computed and further refined by (ii) MCMC simulations. These two sub-steps represent the key algorithmic contribution of VIA.

Lazy-teleporting random walk: We first compute the pseudotime as the expected hitting time of a lazy-teleporting random walk on an undirected cluster-graph generated in Step 1. At each step, the walk (with small probability) can remain or teleport to *any* other state in the graph. The lazy-teleporting nature of this random walk ensures that as the sample size grows, the expected hitting time of each node does not converge to the stationary probability given by local node properties, but instead continues to incorporate the wider global neighborhood.

The cluster graph constructed in VIA is defined as a weighted connected graph **G** (*V*, *E*, *W*) with a vertex set *V* of *n* vertices (or nodes), i.e., $V = \{v_1, \cdots, v_n\}$ and an edge set *E*, i.e., a set of ordered pairs of distinct nodes. *W* is an $n \times n$ weight matrix that describes a set of edge weights between node *i* and *j*, $w_{ij} \geq 0$ are assigned to the edges (v_i, v_j). For an undirected graph, $w_{ij} = w_{ji}$, the $n \times n$ probability transition matrix, *P,* of a standard random walk on **G** is given by

$$P = D^{-1}W \tag{3.1}$$

where $\boldsymbol{D}$ is the $n \times n$ degree matrix, which is a diagonal matrix of the weighted sum of the degree of each node, i.e., the matrix elements are expressed as

$$d_{ij} = \sum_k w_{ik}, i = j \text{ and } 0 \text{ for } \quad i \neq j \tag{3.2}$$

where *k* are the neighbouring nodes connected to node *i*. Hence, d_{ii}(which can be reduced as d_i) is the degree of node *i*. We next consider a *lazy* random walk, defined as *Z*, with probability $(1 - x)$ of being lazy (where $0 < x < 1$), i.e., staying at the same node, then

$$Z = xP + (1 - x)I \tag{3.3}$$

where *I* is the identity matrix. When teleportation occurs with a probability $(1 - \alpha)$, the modified lazy-teleporting random walk *Z′* can be written as follows, where *J* is an $n \times n$ matrix of ones.

$$Z' = \alpha Z + (1 - \alpha)\frac{1}{n}J$$

The hitting time expression is a function of the Green's function (the inverse of the Laplacian associated with the random walk). In our case of a lazy-teleporting random walk, we derive a modified version of the Green's Function $R_{\beta,NL}$ where $\beta = \frac{2(1-\alpha)}{(2-\alpha)}$, and $R_{\beta,NL} = \sum_{m=1} \frac{\Phi_m \Phi_m^T}{[\beta + 2x(1-\beta)\eta_m]}$.

Φ_m and η_m are the mth eigenvector and eigenvalue of the normalized Laplacian (NL).

In the expression of $R_{\beta,NL}$, the β (teleportation) and *x* (laziness) factors regulate the weights given to each eigenvector-value pair in the expected hitting time formulation such that the stationary distribution (corresponding to the term given by the first eigenvalue-vector pair with values 1 and zero, and equal to a measure of the local-node degree-properties) does not overwhelm the global information provided by other "eigenpairs" but can still exercise some influence on the final hitting times. Since the computations are performed on a cluster-graph and not the single-cell level, we can easily incorporate all eigenvalue-eigenvector pairs without causing a bottleneck in runtime. Consequently the modified walk in VIA enables scalable pseudotime computation for large datasets in terms of runtime, but also preserves information about the global neighborhood relationships within the graph.

MCMC Simulations: In the second stage of Step 2, VIA infers the directionality of the graph by biasing the edge-weights with the initial pseudotime computations, and refines the pseudotime through lazy-teleporting MCMC simulations on the forward biased graph. Instead of pruning edges in the "reverse" direction, edge-weights are forward biased based on the time difference between nodes using the logistic function with growth factor b=1.

$$f(t) = \frac{1}{1+e^{-b(t_0-t_1)}} \tag{3.4}$$

This is a form of soft edge thresholding so that we do not preclude transitions in the reverse direction and is a key feature in VIA towards reducing constraints on the random walk. It has since been adopted by methods like CellRank [Lange 2022]. This approach exhibits a greater degree of flexibility in modeling the biology as transitions are not necessarily unidirectional, and also ensures that the graph remains well connected and traversable. We then recompute the pseudotimes on the forward biased graph: Since there is no closed form solution of hitting times on a directed graph, we perform MCMC simulations to determine the pseudotime of nodes (the distance from root to respective nodes). This refinement step ensures that the pseudotime is robust to spurious links (or conversely, links that are too weakly weighted) that can distort calculations based purely on the closed form solution of hitting times.

3.2.4 Automated detection of terminal cell fates and lineage pathways

The algorithm uses the refined directed and weighted graph (edges are re-weighted using the refined pseudotimes) to predict which nodes represent the terminal states based on a consensus vote of pseudotime and multiple vertex connectivity properties, including out-degree (i.e., the number of edges directed out of a node), closeness C(q), and betweenness B(q).

$$C(q) = \frac{1}{\sum_{q \neq r} l(q,r)} \tag{3.5}$$

$$B(q) = \sum_{r \neq q \neq t} \frac{\sigma_{rt}(q)}{\sigma_{rt}} \tag{3.6}$$

VIA then identifies the most likely path of each lineage (from root to terminal state cells) by computing the likelihood of a node traversing towards a particular terminal state (e.g., differentiation). These lineage likelihoods are computed as the visitation frequency under lazy-teleporting MCMC simulations from the root to a particular terminal state, i.e., the probability of node i reaching terminal-state j as the number of times cell i is visited along a successful path (i.e., terminal-state j is reached) divided by the number of times cell i is visited along all of the simulations. The single-cell level KNN graph constructed in Step 1 is then used to project the lineage probabilities of trajectories (pathways from root to cell fate), and temporal ordering derived from the cluster-graph topology onto a single-cell level. In contrast to other trajectory reconstruction methods which compute the shortest paths between root and terminal node [Street et al., 2018, Setty et al., 2019], the lazy-teleporting MCMC simulations in VIA offer a probabilistic view of pathways under relaxed conditions that are not only restricted to the random-walk along a unidirectional tree-like graph, but can also be generalizable to other types of topologies, such as cyclic or connected/disconnected paths.

3.2.5 Downstream visualization of lineage pathways and gene dynamics

Together, these four steps facilitate holistic topological visualization of TI on the single-cell level (e.g., using UMAP [McInnes et al., 2018] or PHATE [Moon et al., 2019]) and enable data-driven downstream analyses such as recovering gene expression trends and single-cell level pathways of lineages, that are essential to biological validation and discovery of lineage commitment. For example, VIA automatically draws temporal gene dynamics of different lineages by using General Additive Models (GAMs) which weight the contribution of a cell towards the expression of a gene along a given point in pseudotime by the single-cell lineage probability of that cell towards that lineage.

3.3 VIA results

We benchmarked VIA on synthetic and real biological data. The main purpose of evaluation on synthetic datasets was to objectively test the ability to capture more complex non-tree like trajectories and quantify the results against a known reference model. The real datasets highlight the ability of VIA to accurately detect cell fates (a result of the flexible probabilistic graph network underlying the TI), scale to large datasets whilst preserving global neighborhood information, and be easily applied to both transcriptomic and non-transcriptomic single cell data.

3.3.1 Simulated data with complex topologies

The availability of a “ground truth” topology in the synthetic datasets with clearly labeled cell fates and time-stamps for each cell also allowed us to quantify different aspects of the trajectory inference using metrics like the Graph Edit Distance, F1-Branch Score (which are both good measures of local similarity between the reference and inferred topologies) and the Ipsen-Mikaihlov metric, and Pearson correlation of the inference and reference pseudotimes.

The differences in accuracy between VIA and other methods is most significant for complex topologies, particularly those with disconnected components comprising cyclic elements.

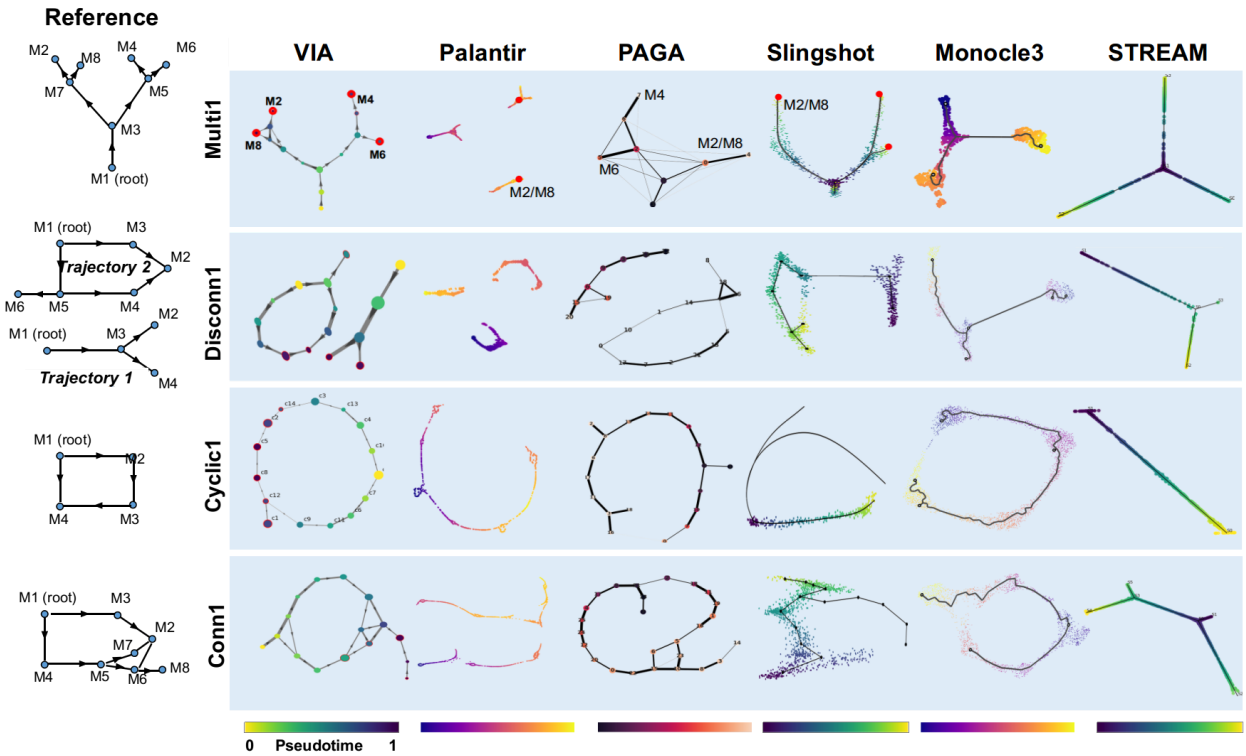


Figure 3.2 TI performance comparisons on complex hybrid topologies. Topologies of four representative synthetic datasets (Multifurc1, Cyclic1, Disconn1, and Conn1) output by different TI methods. The reference topologies are shown on the left. Each dataset contains 1000 "cells" and is run with ten PCs and KNN = 20. VIA is shown at the cluster graph level but can also be projected to the single-cell level as shown in later examples.

3.3.2 Detection of elusive cell fates in endocrine genesis

A detailed analysis of 9 biological datasets can be found in the publication. Here we highlight key findings in 2 of these datasets to provide examples of how VIA overcomes some of the challenges outlined above. A third imaging cytometry dataset of the cell cycle will be introduced in Chapter 4 which is dedicated to computational analyses on image based morphological features of single cell data.

The first dataset we highlight here is the pancreatic dataset of E15.5 murine pancreatic cells spanning developmental stages from initial endocrine progenitor-precursor (EP) state (low level of Ngn3), to intermediate EP (high level of Ngn3) and Fev+ states, to terminal states of hormone-producing alpha, beta, epsilon and delta cells [Bastidas-Ponce et al., 2019] (Fig. 3.3a). A key challenge in analyzing this dataset is the automated detection of the small delta-cell population (a mere 3% of the total population, which requires manual assignment in other methods that predict cell fates such as CellRank and Palantir). In contrast, the well delineated nodes of the VIA cluster-graph (a result of the accurate terminal state prediction enabled by the lazy-teleporting MCMC property of VIA on the inferred topology) lends itself to automatically detecting this small population of delta cells, together with all other key lineages (alpha, beta and epsilon lineages) (Fig. 3.3a-c). As evidenced by the corresponding gene-expression trend analysis, VIA detects all of the hormone-producing cells including delta cells which show exclusively elevated Hhex, Sst, and Cd24a (Fig. 3.3a–e). To show that this is not a co-incidence of parameter choice, we verify that these populations can be identified for a wide range of chosen highly variable genes (HVGs prior to PCA) and number of PCs (see Supplementary Fig. 1c of Stassen 2021). Interestingly, consistent

with an observation by Bastidas-Ponce et al., we see two distinct groups of Fev+ populations branching away from the Ngn+ populations, one with a tendency towards the Beta islets and the other towards the Alpha islets. VIA also finds two Beta cell fates subpopulations (Beta-1 and Beta-2) (Fig. 3.3b–f) that express common Beta-cell markers, such as Dlk1, Pdx1, but differ in their expressions of Ins1 and Ins2.

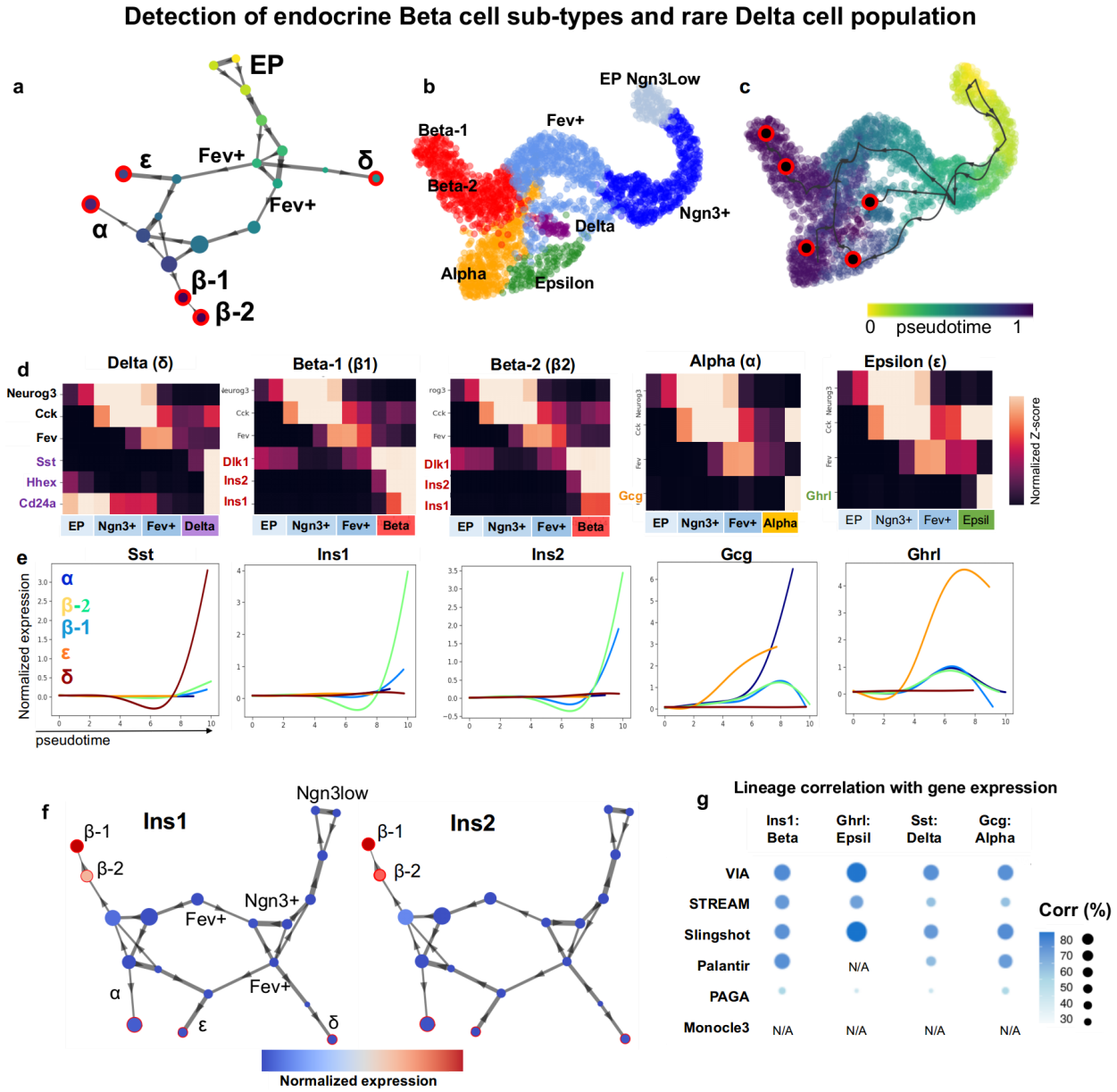


Figure 3.3 Automated detection of islets in endocrine-genesis (a) VIA graph topology Pancreatic Islets: Colored by VIA pseudotime with detected terminal states shown in red and annotated based on known cell type as Alpha, Beta-1, Beta-2, Delta, and Epsilon lineages where Beta-2 is Ins1lowIns2+ Beta subtype. (b) T-SNE colored by reference cell type annotations. (c) colored by inferred pseudotime with predicted cell fates in red-black circles. (d) VIA inferred cluster-level pathway shows gene regulation along endocrine progenitor (EP) to Fev+ cells followed by expression of islet specific genes. (e) Gene-expression trends along pseudotime for each pancreatic islet. (f) Beta-2 subtype expresses Ins2 but not Ins1, suggestive of an immature Beta cell subtype.(g) Marker gene-pseudotime correlations along respective lineages.

3.3.3 Scalability and preservation of global neighborhood information on Mouse Atlas

The 1.3-million scRNA-seq mouse organogenesis cell atlas (MOCA) [Cao 2019] is an important example of VIA's scalability as this dataset is inaccessible to most TI methods from a runtime and memory perspective. VIA can efficiently resolve the underlying developmental heterogeneity, including nine major trajectories (Fig. 3.4) with a runtime of ~40 min. This represents a sizable run time advantage compared to the next fastest method PAGA requiring 3 hours to obtain a coarse graph abstraction excluding any single cell downstream analysis/visualization, and Palantir and STREAM which take over 4 and 6.5 h respectively. Other methods like Slingshot and CellRank were deemed infeasible due to extremely long runtimes on much smaller datasets.

Datasets	Cells	Dimensions after pre-processing	VIA	Palantir	PAGA	Slingshot	CellRank	Monocle3	STREAM
MOCA	1,300,000	30 PCs	40	270	180	X	X	X	390
Mesoderm ESC	89,782	28 proteins	2	6	10	390	NA	NA	15
Endocrine	2,531	30 PCs	1	3	2	3	10	10	1
Human CD34+	5500	200 PCs	2	4	3	150	NA	15	2

Table 3.1 Runtime Comparisons. Computational runtime of VIA and other TI methods (minutes) on four datasets

Beyond the computational efficiency, VIA also preserves wider neighborhood information and reveals a globally connected topology of the MOCA. In contrast, the Monocle3 analysis [Cao et al., 2019] which first reduces the input data dimensionality using UMAP down to 2 dimensions results in several fragmented trajectories that primarily capture the cell type identity of cells (local neighborhood information) and lose the temporal connectivity between different fragments (global information). The overall cluster graph of VIA consists of three main branches that concur with the known developmental process at early organogenesis [Tam 1997] (Fig. 3.4a). It starts from the root stem which has a high concentration of E9.5 early epithelial cells made of multiple sub-trajectories (e.g., epidermis, and foregut/hindgut epithelial cells derived from the ectoderm and endoderm). The stem is connected to two distinct lineages: (1) mesenchymal cells originated from the mesoderm which arises from interactions between the ectoderm and endoderm [Foley et al., 2019, Yao et al., 2019, Hubmap et al., 2019] and (2) neural tube/crest cells derived from neurulation when the ectoderm folds inwards [Gilbert et al., 2000].

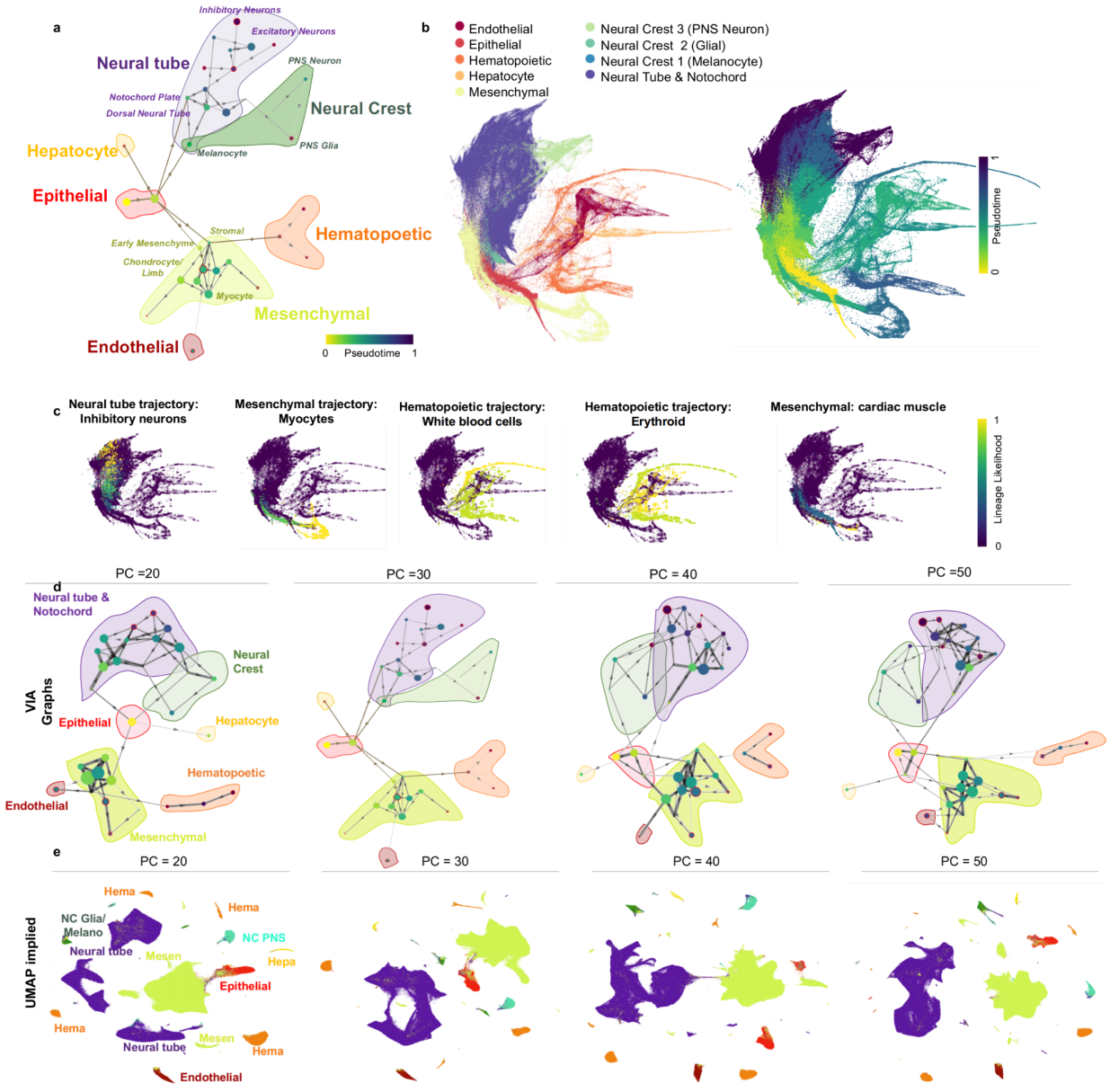


Figure 3.4 Large-scale (1.3 million cells) trajectory inference of mouse organogenesis with VIA (a) MOCA graph trajectory (nodes colored by pseudotime) and shaded-colored regions corresponding to major cell groups. Stem branch consists of epithelial cells derived from ectoderm and endoderm, leading to two main branches: (1) the mesenchymal and (2) the neural tube and neural crest. Other major groups are placed in the biologically relevant neighborhoods, such as the adjacencies between hepatocyte and epithelial trajectories; the neural crest and the neural tube; as well as the links between early mesenchyme with both the hematopoietic cells and the endothelial cells. b Colored by VIA pseudotime. c Lineage pathways and probabilities of neuronal, myocyte and WBC lineages. d VIA graph preserves key relationships across choice of number of PCs, whereas (e) UMAP embedding is first step in Monocle3 and highly susceptible to choice of number of PCs

3.4 Concluding remarks

This section highlighted key challenges to trajectory inference, especially with regards to (i) automatically and accurately predicting terminal cell fates on a variety of single-cell omic datasets, (ii) efficiently handling large-scale single cell datasets and atlases in terms of ensuring reasonable computation times as well as preserving longer range neighborhood connectivity, and (iii) capturing different complex topologies (e.g. linear, tree, cyclic, hybrid). We showed that the cluster-graph based probabilistic approach with soft edge thresholding in VIA improves lineage detection in real datasets at various scales of data size and dimensionality. The computational efficiency also enables faster and more data driven hypothesis testing, which is a valuable trait for biological exploration of new datasets.

VIA has been well received by the single-cell analysis community and is in the top 5% of research outputs as ranked by altmetric both during its time on bioRxiv and now on Nature Communications. It has been adopted as part of various single-cell analysis pipelines either in its entirety (e.g. the Cytograph pipeline from Lugilab for T cell analysis) or adapted in terms of key algorithmic steps to other methods which wrap the VIA computed transition matrix (e.g. soft thresholding concept of edges to allow reverse transitions in CellRank). It is currently being used at AlphaLab HKU to study the progression of Sars-Cov-2 virus infected single cells using image based biophysical features. The download rates on Github and Pypi Stats are also indicators of uptake in the community with 350+ installations per month.

We anticipate the release of VIA 2.0 towards the end of the year, with manuscript preparations currently underway. VIA 2.0 will introduce a “Hybrid TI” approach which allows researchers to augment expression level data (gene or surface marker) together with any available temporal, spatial and RNA-velocity information in order to infer and refine directionality, automate initial state detection. VIA 2.0 will also offer significantly more interpretable and interactive visualizations both on the graph and single-cell manifold level. Some examples of usage and tutorials for the new VIA 2.0 are already available online.

Chapter 4: Analysis of image based data

4.1 Introduction

While scRNA-seq is generally accepted as a gold standard for biological discovery and offers a very comprehensive framework to correlate phenotype and genotype, the purpose of this chapter is to, as a proof of concept, show that biophysical and morphological features of cells extracted based on their single-cell images can also be used to accurately delineate cell types (using clustering and classification methods) as well as predict the sequence of progression in biological developmental processes (using unsupervised trajectory inference methods). Since the use of marker genes or a genetic signature to annotate cell types is an accepted practice, one could propose that an analogous set of biophysical features, captured by optical imaging, may one day also serve as a signatory indicator of cell type.

A substantial body of work already supports the idea that such cellular biophysical properties, extracted from label-free optical imaging [Otto et al., 2015; Kasprowicz et al., 2017], are indeed effective intrinsic markers for probing cellular heterogeneity and cellular processes (e.g. cell proliferation, death, differentiation and malignancy). For example, shapes and textures of cells assessed in bright-field images have been used for the classification of immune cell types [Lippeveld et al., 2019], cell cycle analysis [Blasi et al., 2016] and assigning disease-specific phenotypes for blood analysis [Toepfner et al., 2018]. Quantitative phase imaging (QPI) can be used to quantify cellular dry mass [Park et al., 2018]. Cell mass is primarily composed of intracellular proteins, lipids, metabolites, and nucleic acids [Palm & Thompson 2017] and the regulation of cell mass is thus linked to underlying molecular pathways and single-cell transcriptomic signatures [Kimmerling et al., 2018]. Cell mass related measurements may thus be useful to identify cell types or stages in biological processes.

To contribute to the emerging evidence that biophysical properties (such as cell mass, shape, size, texture, internal spatial arrangement of subcellular components) extracted from images can capture cellular heterogeneity and evolution, we showcase three different types of computational analyses applied to imaging cytometry based data. These are namely clustering (PARC), classification (neural networks) and trajectory inference (VIA). PARC successfully separates digitally mixed lung cancer cell types on the basis of their biophysical image based features without reliance on surface markers (Section 4.2.1). A neural network supervised model combined with transfer learning was also trained to classify lung cancer types across different experimental batches (Section 4.2.2). Finally VIA was applied to snapshot image-based features of a breast cancer cell line undergoing cell cycle progressions to predict the cell cycle sequence based purely on the biophysical features from the QPI images. The fluorescence marker expression related to cell cycle expression was withheld from the unsupervised trajectory inference and only used to validate the inferred progression of cells (Section 4.2.3).

4.2 Results

4.2.1 PARC clusters 1.1 million label-free single-cell images

We use a set of multiple lung cancer cell lines (>1 million cells) to show that cell types can be categorized in both an unsupervised setting (using PARC to cluster the cells in Section 4.2.1) and in a supervised setting (using a Neural network and transfer learning approaches in Section 4.2.2) on the basis of their biophysical attributes derived from label-free single-cell images [Lee et al., 2019a ,Lee et al., 2019b, Siu et al.,2020].

We test the adaptability of PARC to cluster an in-house single-cell image-based dataset which describes the biophysical phenotypic profiles of 1.1 million lung cancer cells across 7 cell lines representing three major subtypes: (i) adenocarcinoma, (ii) squamous cell carcinoma and (iii) small cell carcinoma. The biophysical phenotypes of individual cells were extracted from a recently developed high throughput microfluidic quantitative phase imaging cytometer, multiATOM [Lee 2019a], which captures label-free single-cell images at a high throughput (>10 000 cells/s). In multi-ATOM, each imaged cell generates a Bright Field (BF) image and a Quantitative Phase Image (QPI) from which several biophysical features are derived, for example, cell size, mass (calculated based on the refractive index and cell volume). The clusters produced by PARC unambiguously separate (mean-F1 98.8%) between and within the three broad groups of lung cancer cells (Fig. 4.1). As seen on the heatmap, the three main groups show their characteristic phenotypic profile with subtle differences in some texture features within the same subtype that further differentiate individual cell lines. PARC and Phenograph score the highest in terms of accuracy compared with the other methods (Fig. 4.1), with PARC completing the task in 800 s versus the 7200 for Phenograph using the same computational resources. Seurat is terminated after 5 hours with a memory allocation error (at 120 Gb RAM).

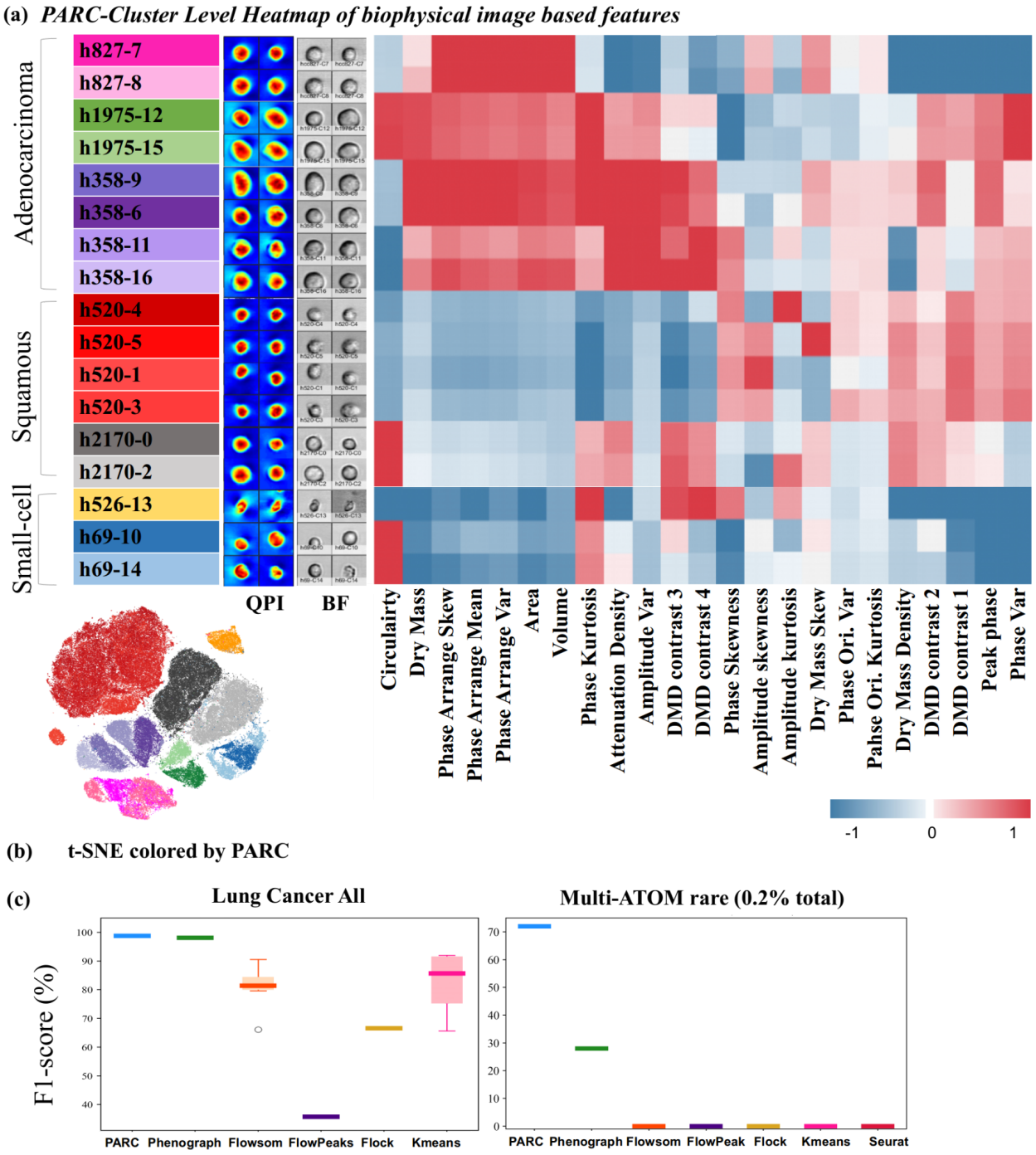


Figure 4.1 Clustering image based data (a) Phenotypic profiles of the cell populations clustered by PARC (with sample QPI and BF images) based on features related to biophysical characteristics extracted from multi-ATOM images. Each of the three main lung cancer subtypes (squamous, adenocarcinoma and small-cell lung cancer) shows its characteristic phenotypic profile. Texture based features further differentiate subtypes within each cell line. Color scale based on Z-score normalized features (b) t-SNE visualization colored by PARC clusters (c) Accuracy of clustering when mixing large populations of each cell type (LHS) and spike test where the rare population is only 0.2% of the full cell population (RHS).

4.2.2 Transfer Learning to classify lung cancer cell types

The specificity of these label-free image based features was further evaluated on the same data with additional experimental batches being added to make up a total population of 2.3 million cells. Specifically, we sought to ask if this label-free method could, across different batches, delineate three major histologically differentiated subtypes of lung cancer amongst seven cell lines, i.e., two subtypes of NSCLC (adenocarcinoma, squamous cell carcinoma) and one for SCLC. We applied a neural network based on the ACTINN architecture [Ma & Pellegrini 2020], combined with transfer learning, to classify the three main subtypes based on the high-dimensional optophysical phenotypic profile as the network inputs (Fig. 4.2). The training dataset is a digitally mixed set of the various cell types whose cell type is known a priori (each cell line was cultured and imaged separately). These training dataset annotations are made available to the neural network during the training stage so that the system 'learns' a biophysical profile associated with each cell line. The testing data consists of cells withheld from the training dataset whose true labels are not provided to the neural network but are known such that the network's performance can be quantified. The features making the phenotypic profile range from bulk features (like cell size, shape and mass) to global and local cellular texture related measurements derived from applying different sized filters/kernels to each image. The 90 features are fed into a neural network consisting of three fully-connected hidden layers of 100, 50 and 25 nodes respectively. A rectified unit function was used as the activation function between them, while the softmax function was used at the output layer, and the Cross-entropy function was selected as a loss function.

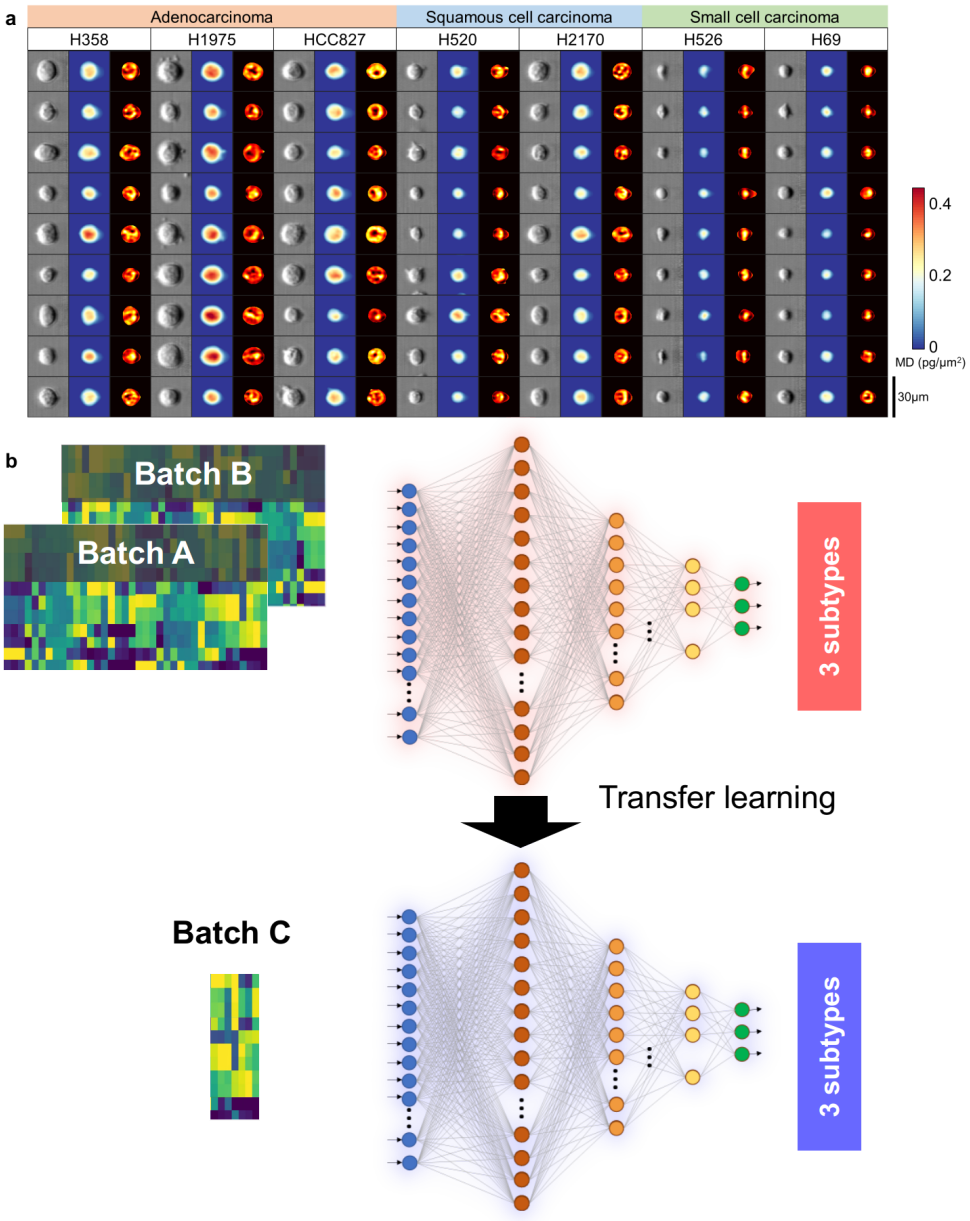


Figure 4.2 Classifying lung cancer cells based on features extracted from imaging cytometry. (a) Randomly selected label-free single-cell images of the seven lung cancer cell lines, which belong to three histologically differentiated subtypes of lung cancer. From left to right in each cell line, the spatial maps are optical density, mass density and local texture transformation of mass density. The colorbar shows the scale of the mass density (MD). (b) General workflow of the label-free lung-cancer subtype classification by transfer-learning-assisted deep neural network. [Adapted from Siu et al., 2020]

To ensure practical optophysical single-cell analysis, we took into account the batch effects in our phenotyping pipeline. Batch effects arise from data variation due to non-biological, technical differences between different repetitions of an experiment and can compromise genuine data interpretation and analysis. Here we adopt a transfer-learning approach that: (1) reduces the training data amount [Yosinski et al., 2014]; (2) and alleviates the batch-to-batch variation problem [Wang et al., 2019]. We found that transfer learning improves the predicted probability for each lung cancer subtype (Fig. 4.3). This was

done by first pre-training the neural network model with large datasets from 2 different batches (A+B). The model parameters were then tuned again by training the model on additional but smaller datasets (5% of the first training dataset) from the third batch (C). Using transfer learning increased the prediction accuracy of the neural network (trained on one batch but tested on another) to a range of 91% to 95%. This is in contrast to the same neural network, which before transfer learning demonstrated significantly lower prediction power. At the same time, the transfer-learning assisted models also demonstrated an improved accuracy ranging (3% to 7%) compared to models being trained and tested on the same batch of data. This improvement could be attributable to the ability to transfer the knowledge of the batch effects (including systematic image focus conditions, system drift, or variations in laser power and photodetector sensitivity) to new classification tasks.

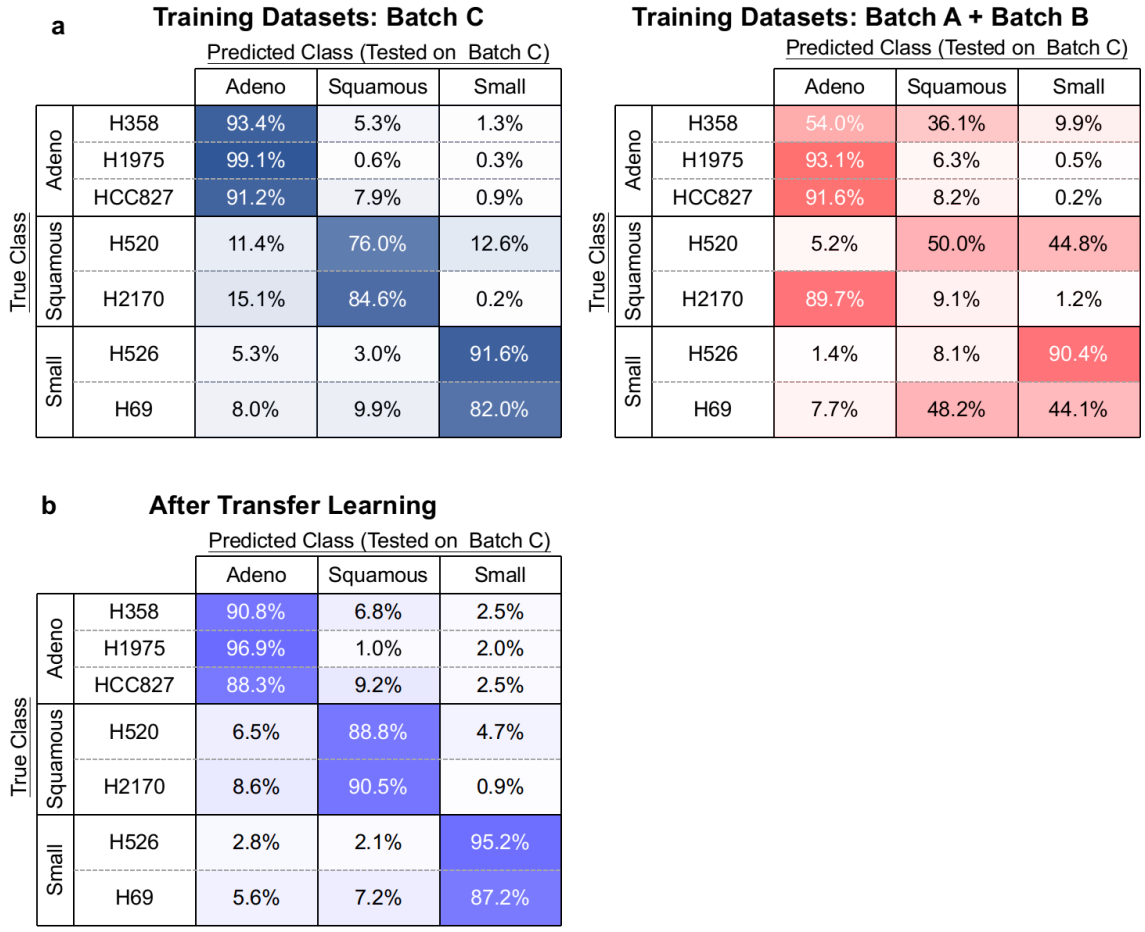


Figure 4.3 Performance for single and multi-batch training, and influence of transfer learning. (a) Confusion matrices (left) of single-batch-trained model (training datasets: Batch C) and (right) of multi-batch-trained model (training datasets: Batch A + B). Models (left) and (right) were trained with 14,000 cells from the training datasets and tested with 105,000 cells from Batch C. The prediction results are reported above. The superior performance of the single-batch-trained model when compared with the others indicates the presence of batch effect. Color gradients of the grids are proportional to the value represented. Darker colors indicate higher values. The highest value in each row is marked with white text. (b) The significance of transfer-learning-assisted neural networks in improving the classification accuracy using label-free optophysical phenotypes (~500 cells from Batch C were used for transfer learning). [Adapted from Siu et al., 2020]

4.2.3 VIA cell cycle trajectory inference

Trajectory predictions based on morphological profiles of single cells have not been studied extensively. Advances in high-throughput imaging cytometry are now making large-scale image data generation and related studies feasible and motivated us to test if VIA can predict biologically relevant progress based on single-cell morphological snapshots captured by a high-throughput imaging flow cytometer. We first generated spatially resolved single-cell biophysical profiles of two live breast cancer cell types (MDA-MB231 and MCF7) undergoing cell cycle progressions (38 features including cell shape, size, dry mass density, optical density and their subcellular textures (Fig. 4.4a). VIA reliably reconstructed the continuous cell-cycle progressions from G1-S-G2/M phase of both types of live breast cancer cells. Intriguingly, according to the pseudotime ordered by VIA (Fig. 4.4c,d), we not only recovered the known cell growth in size and mass [Popescu et al., 2008], and general conservation of cell mass density [Kim et al., 2020] throughout the G1/S/G2 phases, but also a slow-down trend during the G1/S transition in both cell types (Fig. 4.4 f,g), consistent with the lower protein-accumulation rate during S phase [Kafri 2013]. Other methods on this dataset are sensitive to the choice of early cells and detecting intermediate cells as terminal cell fates (e.g. Palantir, Slingshot), and often adding additional edges or branches (e.g., STREAM, PAGA) (See Stassen et al., 2021 Supplementary Fig. S23, S25, S26). The slowdown during the S-phase is missed by the gene trend prediction available in other methods. These results further substantiate the growing body of work [Park et al., 2020, Zangle et al., 2014, Otto et al., 2015] on imaging biophysical cytometry for gaining a mechanistic understanding of biological systems, especially when combined with omics analysis [Kimmerling et al., 2018].

4.3 Concluding remarks

This section showcased three distinct downstream analyses, typically applied to scRNA-seq or mass cytometry based data, that we extended to single-cell image based data. We used graph based unsupervised methods in 1) PARC for clustering of several lung cancer cell types, and 2) VIA for trajectory inference of the cell cycle and identifying accompanying biophysical trends in terms of cell mass and shape of live breast cancer cells and 3) an supervised learning method taking advantage of transfer learning to overcome batch effects in the classification of major lung cancer cell types across multiple batches. These analyses contribute to the emerging body of evidence that cellular heterogeneity can be captured and quantified by these image based phenotypes.

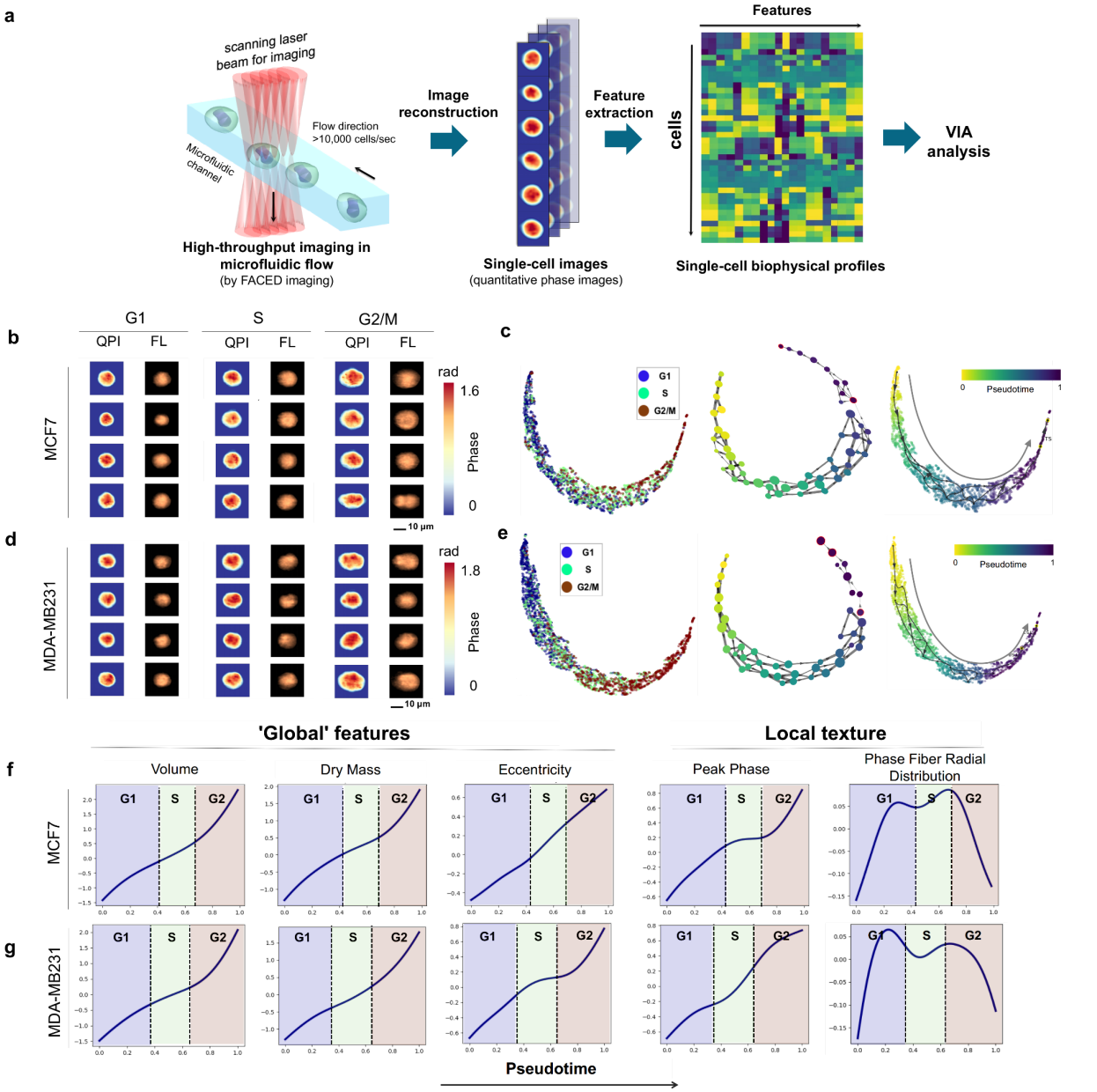


Figure 4.4 VIA infers the cell cycle process using imaging cytometry based features (a) FACED high-throughput imaging flow cytometry of MDA-MB231 and MCF7 cells, followed by image reconstruction and biophysical feature extraction. b Randomly sampled quantitative phase images (QPI) and fluorescence images (FL) of MCF7 cells and (d) MDA-MB231 cells. c Single-cell UMAP embedding colored by the known cell-cycle phase (left), given by DNA-labeled fluorescence images. VIA inferred cluster-graph topology, nodes colored by pseudotime (mid) and UMAP colored by VIA pseudotime for MCF7. d-e VIA analysis repeated for MDA-MB231 cells. f Unsupervised image-feature-trends of global and local biophysical textures against VIA pseudotime for MCF7 and (g) MDA-MB231 cells Cell cycle pseudotime boundaries are defined here as the intersection of the pseudotime probability density functions of each cell cycle stage (annotated based on fluorescence intensity). [Adapted from Stassen et al., 2021]

Chapter 5: Concluding remarks and future work

The covering document began by introducing the full spectrum of a typical downstream single-cell omics analysis pipeline and the computational methods that are typically featured in it. It also highlighted the need for new computational methods to meet the demands of large-scale high-dimensional and noisy datasets. The subsequent chapters proceeded to present two new unsupervised methods (PARC and VIA) designed to improve the existing offering specifically in clustering and trajectory analysis. Here, we briefly highlight the key strengths of each method (PARC and VIA) and then discuss some of their limitations and the scope for future work.

In recent years, the generation of large datasets in the order of millions of cells offers the potential to capture cellular heterogeneity and discover rare cell types. Clustering is often a first step to probing these datasets and we showed that PARC was uniquely suited to identifying small populations due to its hierarchical data-driven pruning step which helps segregate minor populations. A drawback of the 'hard' clustering approach in PARC, where each cell only takes on one label, is that it does not provide a confidence level or probabilistic view of cluster membership. As discussed in the introduction, large scale single cell data can be noisy, sparse (e.g. scRNA-seq, sc-ATAC-seq), or lack sufficient distinguishing features (low number of surface markers in e.g. cytometry) to distinguish cell types. To account for these inherent uncertainties, a statistical view of cluster assignment may offer useful information as to how confidently one can treat the computed cluster memberships, with a view to highlighting boundary cases and pinpointing branching or transition points. PARC's graph based approach could be extended to optionally compute a soft cluster assignment by for example measuring the change in modularity upon reassigning cell membership, or surveying the membership of neighboring cells in the graph.

As mentioned earlier, clustering is often the first downstream analysis step allowing us to discretize and summarize the data. However, for trajectory inference (TI), the connectivity information present at the single-cell level graph before discretization is critical for understanding transitions and relationships between cells as they undergo processes like differentiation. Our trajectory inference method VIA was able to recover complex topologies in differentiation landscapes by combining PARC's cluster level "summarization", with the inter-cellular linkages formed in the original cell-cell graph. While VIA demonstrated strength in tackling challenges related to speed, terminal state detection and recovery of complex topologies, the wide variety of testing datasets also highlighted limitations of the method.

A key feature in VIA was to relax traversal constraints on the inter-cellular transitions that form lineage pathways. VIA does this by probabilistically allowing reversals and teleportations in order to capture disconnected and cyclic topologies that are inaccessible to other methods. However, a major assumption is that at each step of the lineage pathway modeled by a random walk, the cell has no memory of where it was in the prior step. This is likely to be an oversimplification of the underlying biology and it may be meaningful to incorporate memory in these random walks such that a cell's next step is influenced to some extent also by which earlier pluripotent cell type it originated from in its previous state. This idea could be computationally explored using second order random walks. These types of modified second order walks may also offer additional ways to emphasize long range cell-cell interactions and mitigate the loss of global neighborhood relationships when datasets grow in sample size.

Validation of accuracy without a 'gold standard' reference graph or ground truth generally presents an issue for most unsupervised methods. While several gold standard annotated real datasets were available for clustering benchmarking, it was difficult to find real datasets with reference topologies/edges for TI benchmarking. Instead, we had to rely on synthetic datasets (similar to those used by Saelens in their 2019 extensive benchmarking study of TI methods) to quantify the accuracy of true and inferred graph edges for complex topologies. Although reference topologies were not available for the real datasets, many of the processes (e.g. hematopoiesis) are well studied and the general connectivity and chronology inferred by VIA, as well as the predicted regulation of marker genes, was consistent with the known biology.

Another limitation of VIA is that root (initial) cells can only be automatically predicted when RNA velocity is available. A potential improvement to this could be to use a scoring framework such as CytoTrace [Gulati et al., 2020] which uses gene sets that indicate 'stemness' of cells combined with the number of expressed genes as an indicator of development potential in order to predict initial states.

Offering a visually interpretable summary of the trajectory inference is also a challenge. VIA and other TI methods rely on embeddings like UMAP [McInnes et al., 2018] to project the single-cell inferred pathway even though these manifolds are computed independently of the inferred trajectory and pseudotime. VIA's visualization of the cluster graph topology partially addresses this by using a force-directed layout of the underlying cluster-graph to offer useful visual cues into the topology and directionality of the cellular progression (with PAGA being one of the few other TI methods that offers this type of visualization). However, when offering a single-cell view, VIA, like other TI methods, still relies on projecting results onto externally computed embeddings like UMAP, t-SNE or PHATE [Moon et al., 2019]. As a way to improve the visual interpretability and compatibility of the 2D embedding and the superimposed trajectory, we are exploring how to use the modified, reweighted, redirected single-cell 'via graph', as the input for non-linear dimensionality reduction without incurring computational bottlenecks. In fact, these re-weighted, re-directed graphs (which become the input to embedding methods), can be further augmented using available sequential information like temporal annotations in time-series data.

Finally, although VIA is generalizable to data from different experimental modalities, it does not yet offer a way to integrate multiple datasets across different feature spaces (from different experimental modalities) or stitch datasets taken at different points in time. Offering a graph based data-integration method that is part of the trajectory inference pipeline could be useful for time series data as well as the current emergence of spatial omics datasets combined with scRNA-seq.

Identifying these limitations is the first step towards formulating strategies that can alleviate these challenges, and hence also sets the stage for future work. As the diversity of datasets increases in terms of types of experimental modalities, types of organisms, processes being studied, and magnitude of feature and sample size, we need computational methods to be nimble in their ability to evolve and meet rising requirements.

References

Saelens, W. et al. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* 37, 547–554 (2019).

McInnes, L et al. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* 3, 861 (2018).

Moon, K. R. et al. Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* 37, 1482–1492 (2019).

Jia, G. et al. Single cell RNA-seq and ATAC-seq analysis of cardiac progenitor cell transition states and lineage settlement. *Nat. Commun.* 9, 4877 (2018).

Bastidas-Ponce A et al. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrine genesis. Development. 146(12) (2019)

Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genom.* 19, 477 (2018).

Setty, M. et al. Characterization of cell fate probabilities in single-cell data with Palantir [published correction appears in Nat Biotechnol.37(10):1237]. *Nat. Biotechnol.* 37, 451–460 (2019).

Chen, H. et al. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat. Commun.* 10, 1903 (2019).

Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 566, 496–502 (2019).

Lee, J., Hyeon, D.Y. & Hwang, D. Single-cell multiomics: technologies and data analysis methods. *Exp Mol Med* 52, 1428–1442 (2020).

Aviv Regev et al. Human Cell Atlas Meeting Participants Science Forum: The Human Cell Atlas eLife 6:e27041. (2017)

Duò A. et al. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, 7, 1141. (2018)

The Tabula Muris Consortium. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* 583, 590–595 (2020)

Ge Y, Sealfon SC. flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. Bioinformatics. (2012)

Gulati GS et al. Single-cell transcriptional diversity is a hallmark of developmental potential. Science. 24;367(6476):405-411. doi: 10.1126/science.aax0249. PMID: 31974247; PMCID: PMC7694873 (2020)

Levine JH et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell. 2015 Jul 2;162(1):184-97. doi: 10.1016/j.cell.2015.05.047. (2015)

M. Lippeveld, C. et al. Classification of Human White Blood Cells Using Machine Learning for Stain-Free Imaging Flow Cytometry, Cytometry, Part A, 2019

T. Blasi et al. Label-free cell cycle analysis for high-throughput imaging flow cytometry, Nat. Comm., (2016)

N. Toepfner et al., Detection of human disease conditions by single-cell morphological phenotyping of blood, Elife, 7, e29213 (2018).

Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9, 5233 (2019).

Luecken, M. D., & Theis, F. J. Current best practices in single‐cell RNA‐seq analysis: a tutorial. Molecular systems biology, 15(6), e8746. (2019).

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 10008, 6, (2008).

Foley, T. E., Hess, B., Savory, J. G. A., Ringuette, R. & Lohnes, D. Role of Cdx factors in early mesodermal fate decisions. Development 146, dev170498 (2019).

Yao, Y., Yao, J. & Boström, K. I. SOX transcription factors in endothelial differentiation and endothelial-mesenchymal transitions. *Front. Cardiovasc. Med.* 6, 30 (2019).

Gilbert, S. F. Developmental Biology. 6th edn. (Sinauer Associates). The Neural Crest. https://www.ncbi.nlm.nih.gov/books/NBK10065/ (2000).

Park, S. R. et al. Single-cell transcriptome analysis of colon cancer cell response to 5-fluorouracil-induced DNA damage. *Cell Rep.* 32, 108077 (2020).

Tam, P. P. & Behringer, R. R. Mouse gastrulation: the formation of a mammalian body plan. Mech. Dev. 68, 3–25 (1997).

Y. Park, C. Depeursinge and G. Popescu, Quantitative phase imaging in biomedicine, Nat. Photonics, 2018, 12(10), 578–589.

Zangle, T. A. & Teitell, M. A. Live-cell mass profiling: an emerging approach in quantitative biophysics. *Nat. Methods* 11, 1221–1228 (2014).

Mir, M. Optical measurement of cycle-dependent cell growth, Proc. Natl. Acad. Sci. U. S. A., 2011, 108(32), 13124–13129.

R. J. Ellis, Macromolecular crowding: obvious but underappreciated, Trends Biochem. Sci., 26(10) (2001)

M. Al-Habori, Macromolecular crowding and its role as intracellular signalling of cell volume regulation, Int. J. Biochem. Cell Biol., 33(9), 844–864.(2001)

W. Palm and C. B. Thompson, Nutrient acquisition strategies of mammalian cells, Nature, 546(7657), 234–242. (2017)

Tse, H. T. et al. Quantitative diagnosis of malignant pleural effusions by single-cell mechanophenotyping. *Sci. Transl. Med.* 5, 212ra163 (2013).

Otto, O. et al. Real-time deformability cytometry: on-the-fly cell mechanical phenotyping. *Nat. Methods* 12, 199–202 (2015).

Popescu, G. et al. Optical imaging of cell mass and growth dynamics. *Am. J. Physiol. Cell Physiol.* 295, C538–C544 (2008).

Kim, K. & Guck, J. The relative densities of cytoplasm and nuclear compartments are robust against strong perturbation. Biophys. J. 119, 1946–1957 (2020).

Kafri, R. et al. Dynamics extracted from fixed cells reveal feedback linking cell growth to cell cycle. Nature 494, 480–483 (2013)

Kimmerling, R. J. et al. Linking single-cell measurements of mass, growth rate, and gene expression. *Genome Biol.* 19, 207 (2018).

Siu, K. C. M. et al. Deep-learning-assisted biophysical imaging cytometry at massive throughput delineates cell population heterogeneity. *Lab. Chip.* 20, 3696–3708 (2020).

J. Yosinski, J. Clune, Y. Bengio and H. Lipson, How transferable are features in deep neural networks?, Advances in neural information processing systems, pp. 3320–3328. (2014)

Wang T et al. BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes, Genome Biol., 20(1), 1–15. (2019)

F. Ma and M. Pellegrini, ACTINN: automated identification of cell types in single cell RNA sequencing, Bioinformatics, 36(2), 533–538. (2020)

Zheng G et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun*. (2017)

Lee K.C.M. et al. Quantitative phase imaging flow cytometry for ultra‐large‐scale single‐cell biophysical phenotyping. *Cytometry A*, 95, 510–520. (2019b)

Lee K.C.M. et al. Multi‐ATOM: ultrahigh‐throughput single‐cell quantitative phase imaging with subcellular resolution. *J. Biophotonics*, 12, e201800479. (2019a)

Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 566 (2019).

Kasprowicz R. et al. Characterising live cell behaviour: Traditional label-free and quantitative phase imaging approaches. *Int. J. Biochem. Cell Biol.*, 84, 89–95. (2017)

Qiu, C., Cao, J., Martin, B.K. *et al.* Systematic reconstruction of cellular trajectories across mouse embryogenesis. *Nat Genet* 54, 328–341 (2022)

Ong S.M. et al. The pro-inflammatory phenotype of the human non-classical monocyte subset is attributed to senescence. Cell Death Dis., 9, 266. (2018)

Chen, Z., Goldwasser, J., Tuckman, P. et al. Forest Fire Clustering for single-cell sequencing combines iterative label propagation with parallelized Monte Carlo simulations. Nat Commun 13, 3538 (2022).

Weber & Robinson. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. Cytometry Part A. 89. 10.1002/cyto.a.2303 (2016).

Stassen et al. Generalized and scalable trajectory inference in single-cell omics data with VIA. Nat Commun 12, 5528 (2021).

Stassen et al., PARC: ultrafast and accurate clustering of phenotypic data of millions of single cells, Bioinformatics, Volume 36, Issue 9 (2020).

Appendix: Bibliography of works by Shobana V Stassen

Publications:

Shobana V. Stassen, Yip, G.G.K., Wong, K.K.Y. et al. Generalized and scalable trajectory inference in single-cell omics data with VIA. Nat Commun 12, 5528 (2021). https://doi.org/10.1038/s41467-021-25773-3

Shobana V. Stassen, Dickson M. D. Siu, Kelvin C. M. Lee, Joshua W. K. Ho, Hayden K. H. So, Kevin K. Tsia. "PARC: ultrafast and accurate clustering of phenotypic data of millions of single cells". Bioinformatics. 36(9): 2778-2786 (2020).

Siu, Dickson, Lee Kelvin, Lo Michelle, **Shobana V. Stassen,** Wang Maolin, Zhang Iris, So Hayden, Chan Godfrey, Cheah Kathryn, Wong Kenneth, Hsin Michael, Ho James, Tsia Kevin. "Deep-learning-assisted biophysical imaging cytometry at massive throughput delineates cell population heterogeneity." Lab on a chip. (2020)

Conferences

Shobana V. Stassen, Kelvin C. M. Lee, Kevin K. Tsia, "Accelerated Pheno-Tree (APT) for large-scale, label-free of image-based single-cell analysis". High-Speed Biomedical Imaging and Spectroscopy IV 10889, SPIE Photonics West BIOS Speaker (Feb 2019).

Michelle C.K. Lo, **Shobana V. Stassen**, Dickson M.D. Siu, and Kevin K. Tsia, "Robust Quantitative Phase Imaging Cytometry with Transfer Learning". Biophotonics Congress: Biomedical Optics 2020 (Translational, Microscopy, OCT, OTS, BRAIN).

Shobana V. Stassen, K. K. Tsia, "VIA for generalized and scalable single-cell trajectory inference beyond transcriptomic data". CYTO Virtual Interactive 2021 Speaker.

Gwinky Yip, Alex Chin, **Shobana V Stassen,** Michelle CK Lo, Rashmi Sreeramachandramurthy, Kelvin CM Lee, Kenneth KY Wong, Leo LM Poon, Kevin K Tsia, Image-based single-cell biophysical phenotyping of SARS-CoV-2 infection by high-throughput quantitative phase imaging flow cytometry, High-Speed Biomedical Imaging and Spectroscopy VII SPIE 2022.