Single-cell Analysis from the perspective of how to Interact, Identify and Integrate cells

Abdelaal, T.R.M.

DOI
10.4233/uuid:6a9954ba-1a15-4aaa-93f8-b3b49aa55f96

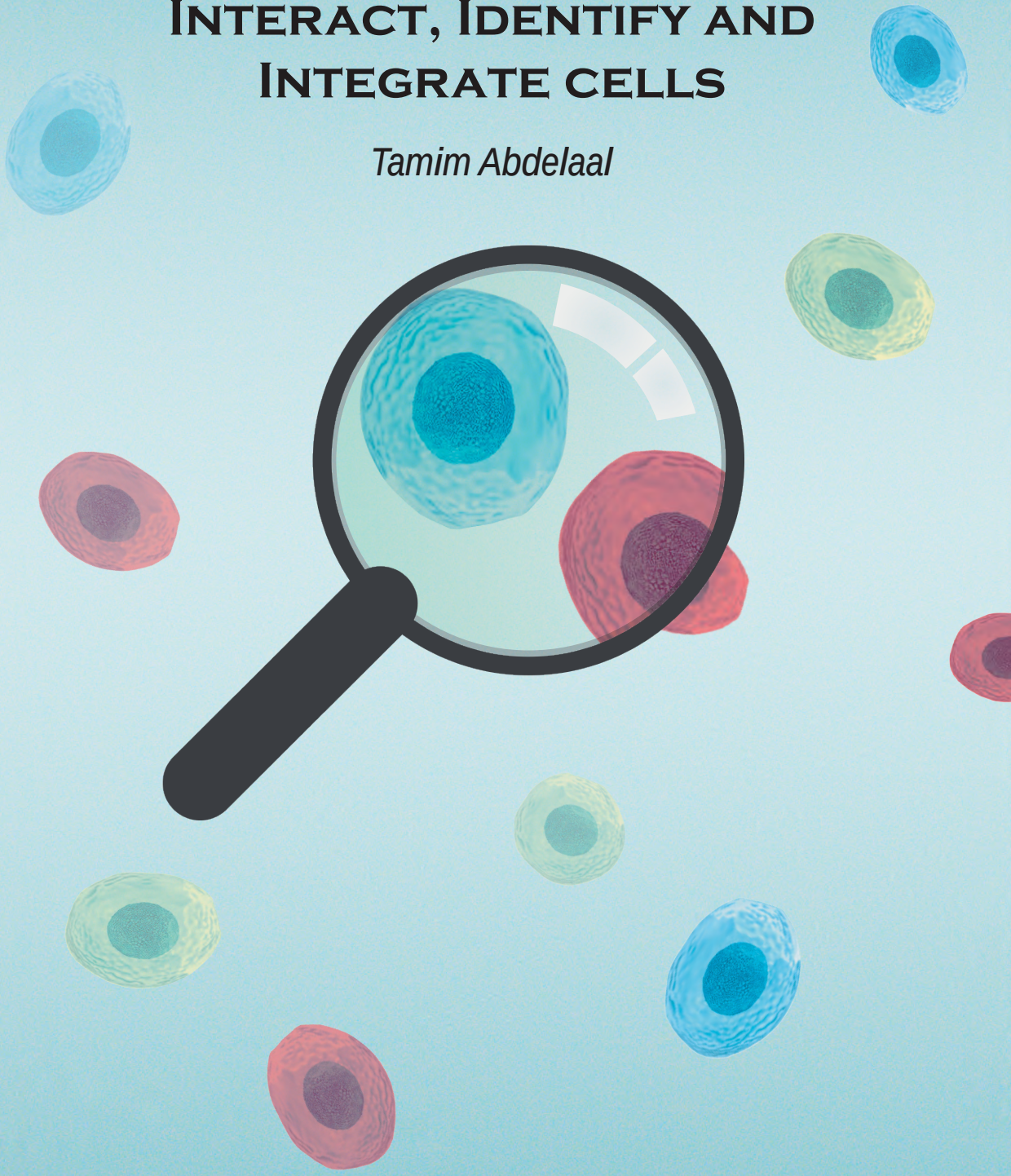
Publication date
2021

Document Version
Final published version

Citation (APA)
Abdelaal, T. R. M. (2021). *Single-cell Analysis from the perspective of how to Interact, Identify and Integrate cells*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:6a9954ba-1a15-4aaa-93f8-b3b49aa55f96

Single-cell Analysis from the perspective of how to Interact, Identify and Integrate cells

Tamim Abdelaal



Propositions

accompanying the dissertation

SINGLE-CELL ANALYSIS FROM THE PERSPECTIVE OF HOW TO INTERACT, IDENTIFY AND INTEGRATE CELLS

by

Tamim Roshdy Mohamed ABDELAAL

1. To scale up single-cell data analysis, new data summarization techniques are needed.

2. In single-cell data, linear models are sufficient for cell type identification (This thesis).

3. In single-cell analysis, extending the number of molecular features through data integration is essential to fully capture cellular heterogeneity (This thesis).

4. For cell type identification in single-cell data, supervised learning can only replace unsupervised learning if a reject option is implemented (This thesis).

5. Researchers should avoid over-interpretation of single-cell data and spend more time on validating their biological findings.

6. Computational developments should not aim to solve short term technological limitations.

7. Every cell is unique.

8. Continuous and comprehensive benchmarking of single-cell analysis methods is essential to eventually reach standardized analysis.

9. A two-dimensional embedding of high-dimensional data should only be used for visualization and interpretation purposes and not for downstream analyses.

10. In the current format, virtual conferences are negatively affecting the scientific community.

These propositions are regarded as opposable and defendable, and have been approved as such by the promotors Prof.dr.ir. M.J.T. Reinders and Dr. A.M.E.T.A. Mahfouz

SINGLE-CELL ANALYSIS FROM THE PERSPECTIVE OF HOW TO INTERACT, IDENTIFY AND INTEGRATE CELLS

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Monday 20 September 2021 at 10:00 o'clock

by

Tamim Roshdy Mohamed ABDELAAL

Master of Science in Biomedical Engineering & Systems
Cairo University, Egypt
born in Cairo, Egypt

This dissertation has been approved by the

Promotor: Prof.dr.ir. M.J.T. Reinders and
Copromotor: Dr. A.M.E.T.A. Mahfouz

Composition of the doctoral committee:

Rector Magnificus,	chairman
Prof.dr.ir. M.J.T. Reinders,	Delft University of Technology, promotor
Dr. A.M.E.T.A Mahfouz,	Leiden University Medical Center, copromotor

Independent members:

Prof.dr.ir. G. Jongbloed	Delft University of Technology
Prof.dr. M. Robinson	University of Zurich, Switzerland
Prof.dr. S. Aerts	KU Leuven, Belguim
Prof.dr. L. Franke	Univeristy Medical Center Groningen, Netherlands
Dr. L. Haghverdi	Max Delbrück Center for Molecular Medicine, Germany

Reserve member:

Prof.dr. L.F.A. Wessels	Delft University of Technology Netherlands Cancer Institute, NKI

Printed by: ProefschriftMaken
Front & Back cover: Donia Kandil and Menna Barakat

ISBN 978-94-6423-384-1

An electronic version of this dissertation is available at
http://repository.tudelft.nl/

Contents

CHAPTER 1

INTRODUCTION

1.1 CELLULAR HETEROGENEITY OF HUMAN TISSUES

The cell represents the smallest unit of life, all living organisms are composed of one or more cells, and cells arise from pre-existing cells. These are the three main principles of the cell theory[1]. Continuous improvements of optical microscopes and magnification technology had large impact on various scientific fields including cell biology. In 1665, with the invention of the compound microscope, the cell was first discovered by Robert Hooke when examining plant tissue. In 1839, Theodor Schwann and Matthias Schleiden confirmed the principle that all animals and plants are composed of cells, and first developed the cell theory.

The number of cells in plants and animals varies per species, the human body contains an estimated of 3.7×10^{13} cells[2], while a complex organ like the human brain is composed of $\sim 80 \times 10^{9}$ cells[3]. All these cells arise from a single-cell through the process of cell division and differentiation, producing a large heterogeneous poll of different cell types across different tissues producing different functions in the human body (e.g. muscle cells, brain cells, liver cells, etc..). Even within one tissue, a heterogeneous cellular composition can still be observed. If we consider blood cells, for example, a hematopoietic stem cell can differentiate into platelets (thrombocytes), red blood cells (erythrocytes) and white blood cells representing the immune system. The immune system is composed of a variety of cell types playing important roles in the innate (e.g. macrophages and neutrophils) and the adaptive (e.g. T and B lymphocytes) immune responses (Figure 1.1A). The cellular composition heterogeneity can be more pronounced when considering the cell state, for example, a naïve T cell may become activated to effector T cell having the ability to kill infected cells, further become a memory T cell producing fast immune response to repeated infection. Thus, studying the cellular composition is crucial to understand the underlying functions and processes within different systems or tissues. In addition to the cellular composition, studying the abundance across different cell types may indicate abnormalities. For example, an elevated abundance of neutrophils in a certain tissue indicates inflammation, while in the pancreas, the decreased abundance of beta cells is a characteristic of type 1 diabetes.

The cellular heterogeneity can be partially resolved based on the cell's morphology. Using optical microscopes, different cell types can be defined based on their shape, size, structure and form. However, cell morphology cannot reveal subpopulations with different cellular states. Studying various cellular molecules helps revealing the cell identity further. For instance, mRNA expression profiles provide a wider view of the cell identity. Fluorescence hybridization techniques can be used to detect specific mRNA molecules analyzed further using fluorescence microscopes[4].

1.2 CELLULAR MOLECULAR FEATURES (BULK VS SINGLE-CELL)

In the 1970s, the first-generation of DNA- and RNA-sequencing methods were introduced, which were able to measure chunks of the genome or the full sequence of some bacterial and viral species with relatively small genomes[5]. In the late 1990s, the second-generation methods, often called next-generation sequencing (NGS), were developed and introduced as high-throughput sequencing methods. NGS technologies are highly scalable to large genomes and often allows the full genome to be measured at once. Over the past 20 years, large quantities of genomic data were generated, where individual RNA and DNA molecules are represented by sequencing reads keeping information on genotypes, phenotypes and cellular states.

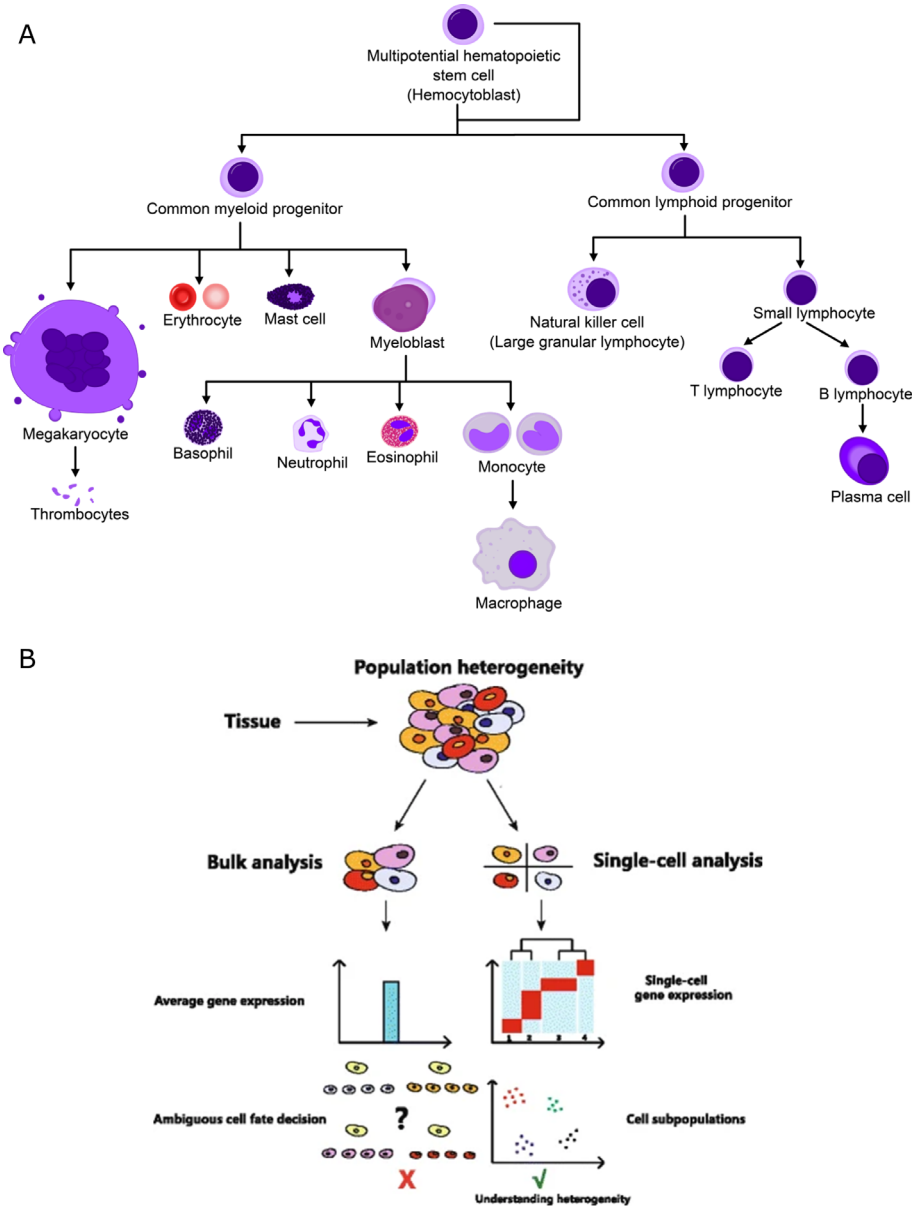


Figure 1.1 (A) Overview of human hematopoiesis [By A. Rad and M. Häggström. CC-BY-SA 3.0 license]. **(B)** Bulk analysis vs single-cell analysis (figure adapted from Fig. 1 in [6]).

Until recently, molecular profiling methods had mostly been applied in bulk, providing a view of an entire sample. Bulk sequencing of RNA from tissues captures the average expression of transcripts across all cells within a tissue, leaving the cellular diversity completely undetected (Figure 1.1B). Single-cell sequencing technologies have emerged as powerful tools to analyze different molecular features at the single-cell resolution, untangling the cellular heterogeneity within a tissue through the detection of different cell populations.

Currently, single-cell technologies can measure several molecular features capturing different aspects of the cellular state. In this thesis, we mainly focused on three features (Figure 1.2): i) protein expression, either lineage marker surface proteins or intracellular signaling proteins, ii) gene expression, measuring the transcribed mRNA across the entire transcriptome, and iii) the spatial context of single-cells within a tissue. However, nowadays other molecular features can be measured at the single-cell resolution, including DNA sequence[7], DNA methylation[8], chromatin accessibility[9] and histone modifications[10], among others.

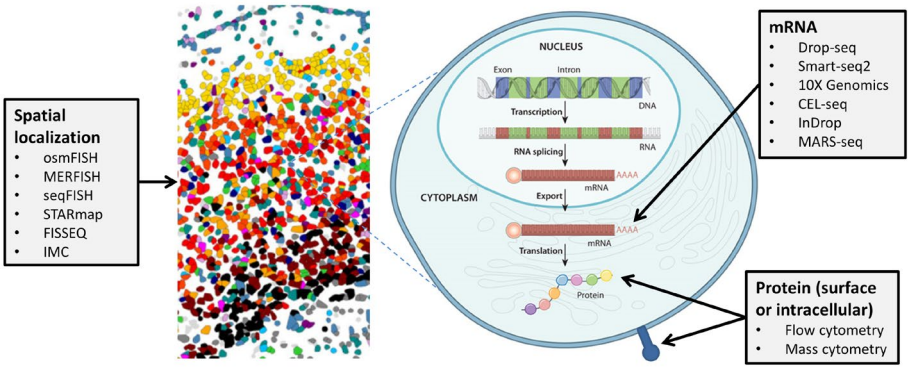


Figure 1.2 Molecular features measured using single-cell technologies.

1.3 TECHNOLOGICAL ADVANCES IN SINGLE-CELL ANALYSIS

Single-cell technologies allow profiling of genomics, transcriptomics and proteomics at the single-cell level across thousands to millions of cells in a single experiment[11]. Flow and mass cytometry can simultaneously detect tens of proteins across millions of cells within the tissue, and single-cell RNA-sequencing technologies enable simultaneous profiling of the full transcriptome across thousands or millions of single-cells[12]. Recently, fluorescence in-situ hybridization and in-situ sequencing technologies can analyze single-cells within tissues and do not require cells to be isolated, thus keeping information about the spatial organization of cells intact[13].

1.3.1 SINGLE-CELL CYTOMETRY FOR PROTEIN MEASUREMENTS

Cytometry is an established single-cell technology for measuring cellular proteins with a high-throughput. In the past few decades, Flow Cytometry (FC) has been the method of choice, in which cells are dissociated and labeled with fluorescent antibodies that bind specific proteins[14]. When activated, these antibodies emit light corresponding with the specific protein abundance in the cell. In addition to the forward- and side-scattered light, the number of antibodies was limited to ~ 14 protein markers due to the light spectra overlap. However, the recently developed full-spectral FC machines are able to simultaneously measure up to 34 markers in a single experiment[15]. FC has been successfully used to characterize different cell types and to isolate specific cell subsets for further analysis (FACS sorting)[16,17]. However, due to the limited number of protein markers, FC cannot be used for a system-wide approach analysis.

Mass Cytometry (CyTOF, cytometry by time-of-flight) overcame the limitation in the number of markers by using heavy metal isotope antibodies[18]. The metal isotopes attached to each

cell are quantified using a time-of-flight mass spectrometer, with which mass indicating different cellular proteins is differentiated. In theory, mass cytometry is capable to simultaneously measure over 100 markers per cell, but practically it's limited to ~50 markers[19]. With the extended number of markers, compared to FC, CyTOF became suitable for system-wide analyses such as profiling the immune system of a specific cancer or tissue. Beside the characterization of canonical cell populations, CyTOF can be used to discover rare (or novel) cell populations. Several studies have illustrated the value of using CyTOF to provide a system-wide view of the cellular composition at the single-cell level, including: studying the immune system heterogeneity of a specific cancer type[20–23], defining disease- or tissue-specific cell populations[24,25], monitoring the immune system response to various infections[26] and immunotherapy[27,28]. However, compared to FC, cells are destroyed during the CyTOF process and cannot be sorted for further analysis.

1.3.2 SINGLE-CELL RNA-SEQUENCING

Single-cell RNA-sequencing (scRNA-seq) has become one of the most widely applied sequencing approaches, providing new opportunities to study and characterize the cellular composition of complex tissues. Typically, cells are first dissociated and isolated, RNA molecules are extracted and transformed into complementary DNA (cDNA) through reverse transcription. Next, cDNA molecules are amplified and sequenced, providing the transcriptome-wide expression of each single-cell. Each single-cell is tagged with a unique DNA barcode that labels the corresponding cDNA molecules. The cDNA molecules from many cells are mixed for sequencing, and transcripts of each cell can be identified by its unique barcode[29]. Rapid and continuous technological advances over the past decade allowed scRNA-seq technologies to exponentially scale-up the number of cells per experiment[30]. Starting from profiling the transcriptome of one single-cell from early embryonic development in 2009[31], reaching a dataset of two million cells from mouse embryos just 10 years later[32]. scRNA-seq protocols differ in their library preparation platforms and the isolation of single-cells, i.e. they can be generally categorized into microwell-based, plate-based and droplet-based. These protocols differ in the number of cells, coverage of the full transcriptome including gene isoforms, and sequencing costs[33].

scRNA-seq has been successfully used to define the cellular composition of complex tissues revealing complex and rare cell populations, discover regulatory relationships between genes, and track the development trajectories of distinct cell lineages. scRNA-seq has been used to study different species including zebrafish[34], frogs[35] and mus musculus[36], as well as different tissues/organs including pancreas[37], peripheral blood mononuclear cells (PBMCs)[38], brain[39], and various types of cancers[22,40,41]. Although most scRNA-seq studies mainly focus on the transcript abundance levels (gene expressions), scRNA-seq may also provide valuable information about the nucleotide sequence such as genetic variants and RNA splicing. In scRNA-seq studies, multiplexing of different donors is a typical experimental setup that can be used to avoid batch effects and reduce sequencing costs. Genetic variants can be detected in the scRNA-seq reads, these variants can be used to de-multiplex and assign the cells to their original donors[42]. Furthermore, scRNA-seq only measures a static snapshot of the transcript abundances in a cell, however, it can also detect the amount of unspliced and spliced RNA. The ratio between unspliced and spliced RNA provides an estimate of the rate of change in transcript abundance, named RNA velocity[43], inferring cellular dynamics which is very important in studying lineage development.

1.3.3 SPATIALLY-RESOLVED SINGLE-CELL DATA

The previously mentioned single-cell technologies, cytometry and scRNA-seq, require cellular dissociation prior to measuring the cells protein or gene expression. As a result, these technologies lack the spatial localization of the cells within a tissue, which limits studying the cellular interaction between different cell populations. Recent advances in the spatial transcriptomics technologies provide gene expression profiles with spatial localization in the tissue. Imaging-based protocols label the mRNA of interest with a fluorescent probe using in situ hybridization (ISH), these techniques produce high gene detection sensitivity but are limited in the number of genes that can be measured simultaneously. Single-molecule fluorescence ISH (smFISH) can simultaneously measure a small number of genes at the single-cell resolution within a tissue[44]. Advanced imaging-based methods such as osmFISH[45], MERFISH[46], seqFISH[47] and seqFISH+[48], can detect up to 10,000 transcripts simultaneously using sequential rounds of hybridization combined with unique barcoding for each transcript. On the other hand, sequencing-based protocols apply in situ RNA sequencing on the tissue. STARmap[49] and FISSEQ[50] can profile a few hundreds to thousands of transcripts, while Spatial Transcriptomics[51] and Slide-seq[52] can profile the whole transcriptome. However, they have a lower cellular resolution and sensitivity to detect genes compared to the imaging-based protocols. Spatial transcriptomics protocols have been widely used, often in combination with scRNA-seq, to study the spatial cellular composition of various organisms and tissues including the drosophila embryo[53], zebrafish embryo[54], and different regions in the mouse brain[45,49,55]. Additionally, they have been used to study the spatial gene expression patterns within tissue sections of various types of cancer including pancreatic[56] and prostate[57] cancer.

Furthermore, Imaging Mass Cytometry (IMC) can analyze single-cell protein expression with spatial localization within a tissue[58]. IMC uses similar principles of the regular CyTOF, where tissue sections are conjugated with protein-specific heavy metal antibodies. The tissue section is ablated using a pulsed laser beam, and the liberated antibody ions are quantified using a time-of-flight mass spectrometer. IMC produces a cellular resolution of 1 μm and can measure up to 40 proteins per tissue. IMC has been mainly used to study cancer tissues and the respective immune system cellular organization[59,60], it has also been successfully used to study the progression of certain diseases such as Type 1 diabetes[61].

1.4 CURRENT PRACTICE IN SINGLE-CELL DATA ANALYSIS

The single-cell experimental workflow requires multiple stages to finally provide a (cells x features) count matrix representing the gene/protein expression patterns of each cell[63,64]. Typical data analysis pipelines include several steps, which can be divided into two major categories: preprocessing and downstream analysis[62] (Figure 1.3). Preprocessing starts with quality control; filtering out low quality cells including debris, doublets and cells with very few detected genes/proteins. Next, normalization is typically applied for scRNA-seq and spatial transcriptomics data, while it is not common for cytometry data, correcting for differences in cell sizes and for variation in mRNA detection across cells. A data transformation usually follows to reduce the skewness of the data and approximate the data to be normally distributed, which is a useful feature for many downstream steps. Feature selection is often applied to select only the informative features for further analysis, e.g. highly variable genes or top principal components using Principal Component Analysis (PCA). Further, preprocessing often involves a batch correction step to correct between technical differences between samples measured on different time points, or samples measured using different machines or protocols.

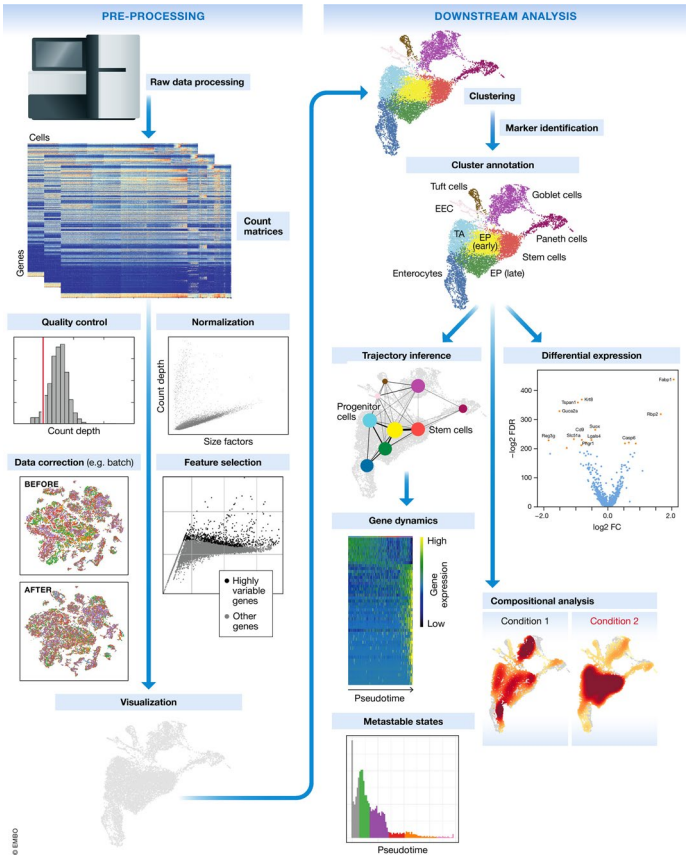


Figure 1.3 Typical single-cell analysis workflow (figure adapted from Figure 1 in [62]).

Downstream analysis mainly starts with the identification of different cell populations within the pool of cells. Cell population identification is usually performed using clustering (unsupervised learning) methods, where cells are grouped into clusters based on the similarity of their gene/protein expression profiles. Next, cell clusters are visualized using low dimensional 2D maps where marker genes/proteins can be overlaid to annotate the clusters with biologically relevant labels. Additionally, further downstream analysis can be carried out including defining differentially expressed genes between cell populations, inferring differentiation trajectories capturing cellular dynamics, studying cell populations distribution across different sample groups (conditions).

1.5 AVENUE FOR IMPROVEMENT IN SINGLE-CELL DATA ANALYSIS

The analysis of single-cell data imposes several challenges due to the large number of cells, as well as the enormous number of measured features per cell. We divide these challenges into three categories (Figure 1.4): i) visual exploration and interactive analysis of the data (*interaction*), ii) definition of the identity of each single-cell (*identification*), and iii) combination of molecular information from different single-cell datasets (*integration*).

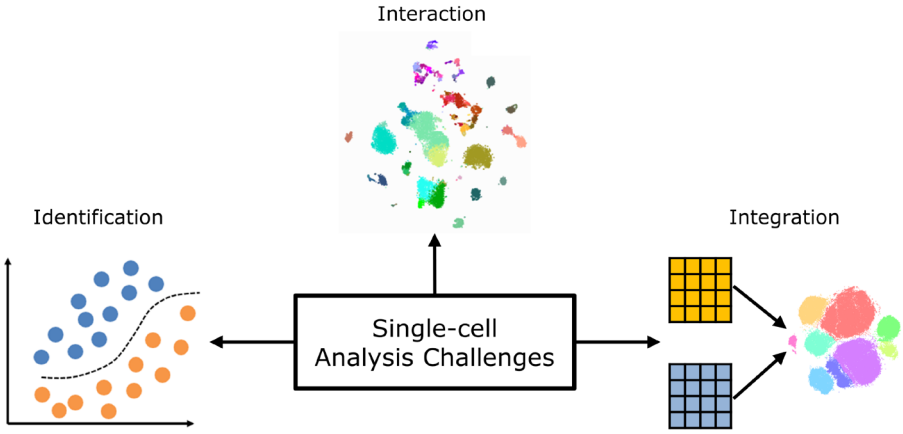


Figure 1.4 Single-cell analysis broad challenges: interaction, identification and integration.

1.5.1 INTERACTION

In single-cell analysis, the identification of different cell populations is a crucial step in the analysis, as it represents the main advantage of single-cell data compared to bulk analysis. Clustering methods are very instrumental in analyzing high-dimensional single-cell data, as they group similar cells. Graph-based community detection methods are the most widely used clustering tools for single-cell data. These methods rely on a graph representation of the data, often obtained using a k-nearest-neighbor approach, which is then partitioned into different clusters. Nowadays, The Louvain[65] community detection algorithm is the default clustering method implemented in the most popular platforms for scRNA-seq analysis including Seurat[66] and Scanpy[67], and in Phenograph[68], a popular clustering method for cytometry data[69]. However, the computational complexity of graph-based clustering increases in a quadratic form with respect to the number of cells, i.e. $O(N^2)$ (Figure 1.5A). Consequently, graph-based clustering is not scalable to current dataset sizes, which typically contain a few millions of cells[30]. Thus, there is a need for robust clustering methods that can scale to millions of cells.

Additionally, after obtaining cell clusters that represent different cell populations, manual input is required to annotate these clusters with biologically relevant labels. This requires visual exploration of the data and inspection of marker genes/proteins expression across different cell populations. This is usually done by overlaying these expressions on a low-dimension representation of the data. Popular non-linear dimensionality reduction such as tSNE[70] and UMAP[71] are often used to visualize the single-cell data. However, also these techniques do not scale well to millions of cells. In addition, from the visualization prospective, tSNE suffers from the "crowding problem" where the resulting two-dimensional map is full of points with no clear distances between different groups of cells, clouding the data structure (Figure 1.5B). Consequently, these limitations prevent the interactive analysis of large datasets having millions of cells.

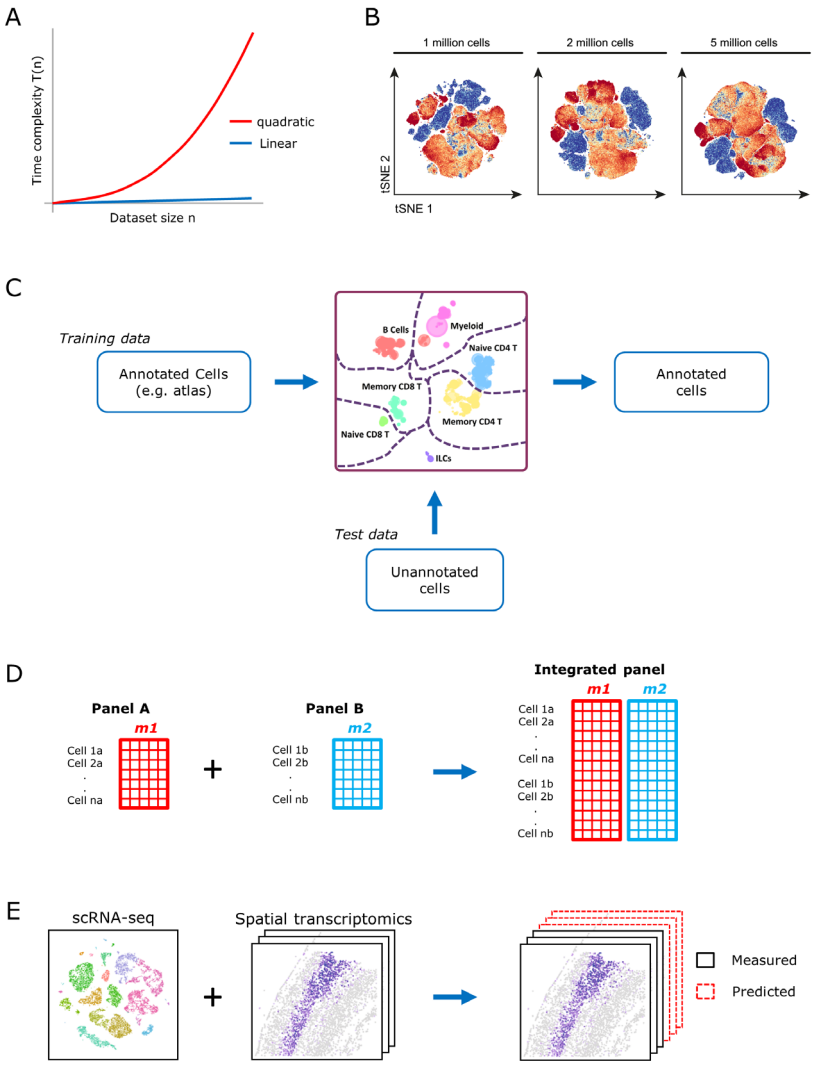


Figure 1.5 Interaction: (A) Computational complexity of linear and quadratic methods, the latter are not scalable to large number of cells. **(B)** tSNE crowding problem, with increased number of cells, distances between different groups of cells become less clear. **Identification: (C)** Classification model can be trained with an annotated dataset to automatically predict cell identities for newly measured cells. **Integration: (D)** Two CyTOF panels, A and B, measuring two sets of protein markers, m1 and m2, can be integrated to extend the number of proteins per cell to m1+m2. **(E)** scRNA-seq, measuring the full transcriptome for dissociated cells, can be integrated with spatial transcriptomics, measuring a limited number of transcripts with spatial localization, to produce a transcriptome-wide spatial gene expression.

1.5.2 IDENTIFICATION

The process of identifying cell population, as described before, works well in explorative experiments, in which all the samples are collected within a short time frame and can be analyzed all at once. However, in large cohort studies with hundreds of samples, the clustering is usually performed per sample, or group of samples, as samples are collected over long time periods, or due to computational limitation in the number of cells that can be

analyzed at once. As a result, the clustering-based annotation becomes tedious and further limits the reproducibility of identifying cell populations across different samples, due to non-deterministic steps in the clustering. Additionally, there is an exponential growth in the number different cell populations that complicates this manual task. These aspects have prompted the development of supervised classification methods for automatic cell identification. These classification methods can be trained using an initial set of samples, or a cell atlas, and next these trained classifiers can be used to automatically predict cell identities for all cells in newly measure samples (Figure 1.5C). Several studies have shown the possibility to use classification for the cell identification task[72–74], however, a comprehensive performance comparison of currently proposed classifiers is lacking, leaving the user unsure which classification method best fits his data. Additionally, it is not clear how complex the classification task is, revealing this complexity would prevent applying over-complicated classification methods, which would improve the classifier interpretability and scalability to larger datasets.

1.5.3 INTEGRATION

Although CyTOF can simultaneously profile ~3x more proteins in a single experiment as compared to flow cytometry, this often is still not enough to capture the full cellular heterogeneity. One way to overcome this is to perform a multi-tube cytometry experiment, in which a sample is divided across multiple tubes which are analyzed using different protein marker panels[75–77]. These different panels can then integrated to extend the protein measurement vector for each cell (Figure 1.5D). This is, however, complicated by the fact that the cells are disassociated with which the matching between cells across the different tubes is lost. Resolving this issue would dramatically increase the power to investigate cellular subpopulations as a much wider protein expression profile is then known for every cell.

Additionally, data integration across different technologies, e.g. integrating scRNA-seq and spatial transcriptomics data, would provide a more complete overview of cellular identities and interactions within complex tissues. scRNA-seq measures the whole-transcriptome expression of dissociated single-cells, lacking their spatial localization. Whereas, spatial transcriptomics technologies do retain valuable information about the spatial context of cells, but are limited in the number of transcripts that can be assessed simultaneously. Integration of scRNA-seq with spatial transcriptomics would give the potential to extend the spatially measured expressions to the full transcriptome (Figure 1.5E). Again, this process is hampered by the disassociation of the cells before scRNA-seq is applied. Being able to match cells between spatial transcriptomics and scRNA-seq would enrich the spatial information about cells enormously.

1.6 THESIS CONTRIBUTIONS

The research presented in this thesis addresses the three challenges identified with a number of computational methods. Regarding the *interaction* challenges, we introduce, in **Chapter 2**, SCHNEL, a clustering method for high dimensional single-cell data which uses in its core the Louvain graph-based clustering. However, SCHNEL is scalable to tens of millions of cells. We show that SCHNEL outperforms state-of-the-art clustering methods and produces reliable clustering in workable time frames. In **Chapter 3**, we introduce Cytosplore-transcriptomics, a complete platform for the interactive analysis of scRNA-seq data. Based on Cytosplore[78,79], previously introduced for interactive analysis of cytometry data having millions of cells and tens of features (proteins), we scale Cytosplore-transcriptomics to scRNA-seq data having

thousands of features (genes). Cytosplore-transcriptomics is capable to apply all mentioned preprocessing steps, interactive data visualization using low-dimensional maps, as well as downstream analysis such as clustering, cell type annotation and detecting differentially expressed genes across cell clusters.

To address the *identification* challenges, we perform, in **Chapter 4**, a comprehensive benchmark evaluating the performance of 22 classification methods to automatically identify cell populations across 27 scRNA-seq datasets of different sizes, species, technologies and level of complexity. Using two experimental cross-validation setups, we show that the majority of the classification methods perform well on a variety of datasets, and that performance decreases when the complexity of the datasets increases, such as overlapping populations or when there are many cell populations. Further, we show that the support vector machine classifier, a general-purpose classifier, has the best performance overall. In **Chapter 5**, we propose the Linear Discriminant Analysis (LDA) classifier for automatic cell identification of cytometry datasets. We show that LDA outperforms previous published methods and show its scalability to large datasets with millions of cells as well as having a large number of cell populations. In **Chapter 6,** we show the potential of the LDA classifier introduced in Chapter 5, by using it to automatically identify cell populations when comparing two cohorts of colorectal cancer patients. With this analysis, we have found a new innate lymphocyte population to be enriched in colorectal cancer tissues, among other immune populations from the innate and adaptive compartments.

Finally, regarding the *integration* challenges, in **Chapter 7** we introduce CyTOFmerge, a data integration platform for CyTOF data. CyTOFmerge integrates protein measurements across multiple marker panels at the single-cell level. These marker panels share a set of common markers serving as the basis for integration. We show that CyTOFmerge outperformed previously introduced integration methods for FC data. Further, we illustrate the benefit of extending the number of protein markers to reveal hidden cell subpopulations. In **Chapter 8**, we introduce SpaGE, a method to predict unmeasured genes for each cell in a spatial transcriptomics dataset through integration with scRNA-seq data from the same tissue. SpaGE relies on domain adaptation to correct for technical differences between both single-cell modalities. SpaGE outperforms state-of-the-art methods in predicting unmeasured spatial gene expression profiles across different regions in the mouse brain, both in accuracy and scalability.

Finally, we conclude the thesis with a discussion of our contributions, potential extensions of our work, together with a brief discussion on the future of single-cell analysis.

BIBLIOGRAPHY

1. Gupta, P. K. *Cell and Molecular Biology*. (Rastogi publications, 2008).
2. Bianconi, E. *et al.* An estimation of the number of cells in the human body. *Ann. Hum. Biol.* **40,** 463–471 (2013).
3. Azevedo, F. A. C. *et al.* Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J. Comp. Neurol.* **513,** 532–541 (2009).
4. Amann, R. & Fuchs, B. M. Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. *Nature Reviews Microbiology* **6,** 339–348 (2008).
5. Onaga, L. A. Ray Wu as Fifth Business: Deconstructing collective memory in the history of DNA sequencing. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **46,** 1–14

(2014).
6. Ye, F., Huang, W. & Guo, G. Studying hematopoiesis using single-cell technologies. *Journal of Hematology and Oncology* **10,** (2017).
7. Vitak, S. A. *et al.* Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* **14,** 302–308 (2017).
8. Mulqueen, R. M. *et al.* Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36,** 428–431 (2018).
9. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523,** 486–490 (2015).
10. Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33,** 1165–1172 (2015).
11. Wang, D. & Bodovitz, S. Single cell analysis: The new frontier in 'omics'. *Trends in Biotechnology* **28,** 281–290 (2010).
12. Wu, A. R., Wang, J., Streets, A. M. & Huang, Y. Single-cell transcriptional analysis. *Annual Review of Analytical Chemistry* **10,** 439–462 (2017).
13. Lee, J. H. De Novo Gene Expression Reconstruction in Space. *Trends in Molecular Medicine* **23,** 583–593 (2017).
14. Picot, J., Guerin, C. L., Le Van Kim, C. & Boulanger, C. M. Flow cytometry: Retrospective, fundamentals and recent instrumentation. *Cytotechnology* **64,** 109–130 (2012).
15. Futamura, K. *et al.* Novel full-spectral flow cytometry with multiple spectrally-adjacent fluorescent proteins and fluorochromes and visualization of in vivo cellular movement. *Cytom. Part A* **87,** 830–842 (2015).
16. McKinnon, K. M. Flow cytometry: An overview. *Curr. Protoc. Immunol.* **2018,** 5.1.1-5.1.11 (2018).
17. Penter, L. *et al.* FACS single cell index sorting is highly reliable and determines immune phenotypes of clonally expanded T cells. *European Journal of Immunology* **48,** 1248–1250 (2018).
18. Bandura, D. R. *et al.* Mass cytometry: Technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* **81,** 6813–6822 (2009).
19. Spitzer, M. H. & Nolan, G. P. Mass Cytometry: Single Cells, Many Features. *Cell* **165,** 780–791 (2016).
20. Chevrier, S. *et al.* An Immune Atlas of Clear Cell Renal Cell Carcinoma. *Cell* **169,** 736–749 (2017).
21. Wagner, J. *et al.* A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. *Cell* **177,** 1330–1345.e18 (2019).
22. Lavin, Y. *et al.* Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses. *Cell* **169,** 750–765 (2017).
23. Zhang, Q. *et al.* Integrated multiomic analysis reveals comprehensive tumour heterogeneity and novel immunophenotypic classification in hepatocellular carcinomas. *Gut* **68,** 2019–2031 (2019).
24. van Unen, V. *et al.* Mass Cytometry of the Human Mucosal Immune System Identifies Tissue- and Disease-Associated Immune Subsets. *Immunity* **44,** 1227–1239 (2016).
25. Wong, M. T. *et al.* A High-Dimensional Atlas of Human T Cell Diversity Reveals Tissue-Specific Trafficking and Cytokine Signatures. *Immunity* **45,** 442–456 (2016).
26. Newell, E. W., Sigal, N., Bendall, S. C., Nolan, G. P. & Davis, M. M. Cytometry by Time-of-Flight Shows Combinatorial Cytokine Expression and Virus-Specific Cell Niches within a Continuum of CD8 + T Cell Phenotypes. *Immunity* **36,** 142–152 (2012).
27. Helmink, B. A. *et al.* B cells and tertiary lymphoid structures promote immunotherapy response. *Nature* **577,** 549–555 (2020).
28. Adams, H. C. *et al.* High-Parameter Mass Cytometry Evaluation of Relapsed/Refractory Multiple Myeloma Patients Treated with Daratumumab Demonstrates Immune Modulation as a Novel Mechanism of Action. *Cytom. Part A* **95,** 279–289 (2019).
29. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells

using nanoliter droplets. *Cell* **161,** 1202–1214 (2015).
30. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols* **13,** 599–604 (2018).
31. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6,** 377–382 (2009).
32. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566,** 496–502 (2019).
33. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine* **50,** (2018).
34. Briggs, J. A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science (80-.).* **360,** (2018).
35. Plass, M. *et al.* Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science (80-.).* **360,** (2018).
36. Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562,** 367–372 (2018).
37. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **3,** 346–360.e4 (2016).
38. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8,** (2017).
39. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563,** 72–78 (2018).
40. Brady, S. W. *et al.* Combating subclonal evolution of resistant cancer phenotypes. *Nat. Commun.* **8,** (2017).
41. Kim, K. T. *et al.* Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. *Genome Biol.* **17,** 80 (2016).
42. Heaton, H. *et al.* Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* **17,** 615–620 (2020).
43. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560,** 494–498 (2018).
44. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5,** 877–879 (2008).
45. Codeluppi, S. *et al.* Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods* **15,** 932–935 (2018).
46. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (80-.).* **348,** (2015).
47. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods* **11,** 360–361 (2014).
48. Eng, C. H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568,** 235–239 (2019).
49. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science (80-.).* **361,** eaat5691 (2018).
50. Lee, J. H. *et al.* Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* **10,** 442–458 (2015).
51. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353,** 78–82 (2016).
52. Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science (80-.).* **363,** 1463–1467 (2019).
53. Karaiskos, N. *et al.* The Drosophila embryo at single-cell transcriptome resolution. *Science (80-.).* **358,** 194–199 (2017).
54. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33,** 495–502 (2015).
55. Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science (80-.).* **362,** (2018).
56. Moncada, R. *et al.* Integrating microarray-based spatial transcriptomics and single-cell

RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat. Biotechnol.* **38,** 333–342 (2020).
57. Berglund, E. *et al.* Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.* **9,** (2018).
58. Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* **11,** 417–422 (2014).
59. Schulz, D. *et al.* Simultaneous Multiplexed Imaging of mRNA and Proteins with Subcellular Resolution in Breast Cancer Tissue Samples by Mass Cytometry. *Cell Syst.* **6,** 25–36.e5 (2018).
60. Ijsselsteijn, M. E., van der Breggen, R., Sarasqueta, A. F., Koning, F. & de Miranda, N. F. C. C. A 40-marker panel for high dimensional characterization of cancer immune microenvironments by imaging mass cytometry. *Front. Immunol.* **10,** (2019).
61. Damond, N. *et al.* A Map of Human Type 1 Diabetes Progression by Imaging Mass Cytometry. *Cell Metab.* **29,** 755–768.e5 (2019).
62. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15,** (2019).
63. Svensson, V. *et al.* Power analysis of single-cell rnA-sequencing experiments. *Nat. Methods* **14,** 381–387 (2017).
64. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* **65,** 631–643.e4 (2017).
65. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008,** (2008).
66. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888–1902.e21 (2019).
67. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19,** (2018).
68. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162,** 184–197 (2015).
69. Weber, L. M. & Robinson, M. D. Comparison of Clustering Methods for High-Dimensional Single-Cell Flow and Mass Cytometry Data. *Cytom. A* **89,** 1084-1–96 (2016).
70. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn.* **9,** 2579–2605 (2008).
71. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37,** 38–47 (2019).
72. Lee, H., Kosoy, R., Becker, C. E., Dudley, J. T. & Kidd, B. A. Automated cell type discovery and classification through knowledge transfer. *Bioinformatics* **33,** 1689–1695 (2017).
73. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15,** 1053–1058 (2018).
74. Kiselev, V. Y., Yiu, A. & Hemberg, M. Scmap: Projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15,** 359–362 (2018).
75. Pedreira, C. E. *et al.* Generation of Flow Cytometry Data Files with a Potentially Infinite Number of Dimensions. *Cytom. A* **73A,** (2008).
76. Lee, G., Finn, W. & Scott, C. Statistical file matching of flow cytometry data. *J. Biomed. Inform.* **44,** 663–676 (2011).
77. O'Neill, K. *et al.* Deep profiling of multitube flow cytometry data. *Bioinformatics* **31,** 1623–1631 (2015).
78. Höllt, T. *et al.* Cytosplore : Interactive Immune Cell Phenotyping for Large Single-Cell Datasets. in *Computer Graphics Forum (Proceedings of EuroVis 2016)* **35,** (2016).
79. Van Unen, V. *et al.* Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. *Nat. Commun.* **8,** 1–10 (2017).

CHAPTER 2

SCHNEL: SCALABLE CLUSTERING OF HIGH DIMENSIONAL SINGLE-CELL DATA

Tamim Abdelaal*

Paul de Raadt*

Boudewijn P.F. Lelieveldt

Marcel J.T. Reinders

Ahmed Mahfouz

This chapter is published in: *Bioinformatics* (2020) 36: i849-i856, doi: 10.1093/bioinformatics/btaa816.
Supplementary material is available online at: https://academic.oup.com/bioinformatics/article/36/Supplement_2/i849/6055909#supplementary-data
*Equal contribution

Single cell data measures multiple cellular markers at the single-cell level for thousands to millions of cells. Identification of distinct cell populations is a key step for further biological understanding, usually performed by clustering this data. Dimensionality reduction based clustering tools are either not scalable to large datasets containing millions of cells, or not fully automated requiring an initial manual estimation of the number of clusters. Graph clustering tools provide automated and reliable clustering for single cell data, but suffer heavily from scalability to large datasets. We developed SCHNEL, a scalable, reliable and automated clustering tool for high-dimensional single-cell data. SCHNEL transforms large high-dimensional data to a hierarchy of datasets containing subsets of data points following the original data manifold. The novel approach of SCHNEL combines this hierarchical representation of the data with graph clustering, making graph clustering scalable to millions of cells. Using seven different cytometry datasets, SCHNEL outperformed three popular clustering tools for cytometry data, and was able to produce meaningful clustering results for datasets of 3.5 and 17.2 million cells within workable timeframes. In addition, we show that SCHNEL is a general clustering tool by applying it to single-cell RNA sequencing data, as well as a popular machine learning benchmark dataset MNIST.

2.1 INTRODUCTION

Cytometry is an established high-throughput technology for measuring cellular proteins at single-cell resolution. In the traditional Flow Cytometry (FC), cells are labeled with fluorescent antibodies that bind specific proteins[1]. Once excited, these antibodies emit light in correspondence with the targeted protein abundance. These light signals limit the number of potential protein markers as the light spectra will eventually overlap. The advanced Mass Cytometry, cytometry by time of flight, or CyTOF expands the number of markers by using metal isotope antibodies[2]. The theoretical upper limit to the number of markers is 100, in practice most CyTOF studies use between 30 and 40 markers[3]. Cytometry, including both FC and CyTOF, has become a vital clinical tool and has been applied to several clinical studies, including, but not limited to: diagnose acute and chronic leukemia[4], monitoring patients' immune systems after hematopoietic stem cell transplantations[5], defining biomarkers in case-control studies[6], and studying the immune cells differentiation in lung cancer[7].

Cytometry is a high-throughput technology resulting in high-dimensional datasets of millions of cells, representing a major challenge in data analysis. A critical step in analyzing cytometry data is grouping the individual cell measurements into distinct cell populations. Traditionally, FC data was manually analyzed by plotting measured intensities of each pair of markers. This allows researchers to gate distinct cell populations by selecting groups of cells with similar protein expression patterns. Cells are grouped by either positive or negative expression of a marker. However, as the number of markers that can be measured increases, the time required for processing this manual labor tremendously increases. This manual gating process is not even applicable for CyTOF data, with ~240 gates that need to be analyzed when using 40 markers. Additionally, manual gating is biased by the person performing the gating and suffers from reproducibility issues. It also assumes dichotomic expression of a marker (either negative or positive), and ignores the potential of a marker possessing a gradient pattern.

Consequently, researchers have turned to computational methods for analyzing cytometry data. Clustering is an unsupervised process of grouping data points (cells) by its features (protein markers) into distinct groups (cell populations). Many tools have already been published for the task of clustering cytometry data into cell populations[8–10]. These tools can be broadly divided into two categories: dimensionality reduction based, and graph based.

In dimensionality reduction approaches, an algorithm first reduces high dimensional data to fewer dimensions in which the data is then clustered. Reducing to two or three dimensions allows visual representation of high dimensional data, which is otherwise impossible. The archetypical dimensionality reduction technique is Principal Component Analysis (PCA). PCA is limited in its usefulness for cytometry data because it fails to capture non-linear patterns which are characteristic of high dimensional omics data. A popular non-linear dimensionality reduction technique in the single cell community is t-distributed stochastic neighbor embedding (tSNE)[11]. tSNE analyses local neighborhoods of data points and tries to embed the shape of the high dimensional data onto a lower number of dimensions. Clustering can then be performed on the low dimensional embedding to reduce the computational burden of clustering in high dimensional space. Tools such as ACCENSE[12], ClusterX[13], and DensVM[14] are all examples of tools that perform clustering after dimensionality reduction. Non-linear dimensionality reduction methods, like tSNE, suffer from scalability to large datasets. Despite recent improvements of the algorithm, calculating tSNE embeddings becomes prohibitively slow for more than million data points[15–17]. Additionally, tSNE embeddings are stochastic and the resulting global structure of the embeddings for identical data will be different between two runs. This can affect any clustering done in the tSNE dimensions; the stochasticity of the embeddings will make the results less reproducible and less reliable.

Hierarchical Stochastic Neighbor Embedding (HSNE) is a machine learning technique that was introduced to solve the scaling problem associated with tSNE. HSNE transforms large volume of high-dimensional data to a hierarchical set of smaller volumes at representing different scales of the data[18,19]. At any scale, the data can be processed, such as making tSNE embeddings to visualize the reduced data and subsequently cluster the data at these scales. HSNE implementations exist in Cytosplore[20] and High Dimensional Inspector (https://github.com/Nicola17/High-Dimensional-Inspector). Cytosplore allows users to cluster the 2D tSNE embeddings of each data scale with Gaussian Mean Shift clustering, remedying the scaling problem as at these scales the volume of the data can be orders of magnitude smaller than the full dataset. Nevertheless, the clustering still suffers from reproducibility and reliability because of the stochastic tSNE step to reduce the dimensionality.

A different dimensionality reduction based tool is FlowSOM, which clusters the data using a self-organized map (SOM)[21]. Briefly, a SOM consists of a grid of nodes, each representing a point in the high-dimensional space. The grid is trained in such a way that closely connected nodes are highly similar. Each point of the dataset is clustered to the most similar node in the grid. FlowSOM does not suffer from scalability issues, as the computation time is extremely fast[10]. However, FlowSOM cannot automatically find the correct number of clusters, producing less accurate clustering when cell populations are more similar.

An alternative to clustering in low dimensional space, is to cluster the data in the original high dimensional space using graph based techniques. Graph clustering tools like Louvain clustering in Phenograph[22] and X-shift[23] start by finding for each data point the k nearest neighbors. The neighborhood graph is then analyzed to find regions with high connectivity, indicating clusters of similar cells. Compared to dimensionality reduction tools, graph clustering tools provide more reproducible, reliable and automated clustering, with a better ability to detect cell populations with relatively few cells. On the other hand, these graph based methods suffers heavily from the scalability to large datasets, exemplified by runtimes for Phenograph and X-shift that exceed 5 hours for a dataset of ~0.5 million cells[10].

Here, we present SCHNEL, a scalable, reliable and automated clustering tool for high-dimensional single-cell data. SCHNEL combines the hierarchy idea of the HSNE transform with a graph based clustering, making graph based clustering scalable to millions of cells.

SCHNEL produces fast and accurate clustering of cytometry datasets, as well as different types of high-dimensional datasets such as the popular machine learning benchmark dataset MNIST and single-cell RNA-sequencing data.

2.2 METHODS

2.2.1 SCHNEL WORKFLOW

We developed SCHNEL, **S**calable **C**lustering of **H**ierarchical stochastic **N**eighbor **E**mbedding hierarchies using **L**ouvain community detection, a novel method for clustering high dimensional data that scales towards millions of cells. It combines the HSNE manifold-preserving data reduction properties with graph clustering to assign each data point to a meaningful cluster, while performing the actual clustering on a reduced subset of the data. It uses the hierarchical information contained in HSNE to assign the predicted cluster labels on a subset of the data, back to the full dataset (Figure 2.1).

2.2.1.1 Creating Hierarchy using HSNE

We used HSNE as introduced by[18] to construct a hierarchical data representation of the entire high-dimensional dataset. In brief, the hierarchy starts with the raw data, which is then aggregated (summarized) to more abstract scales. At the bottom of the hierarchy, the first scale (data scale) S_0 is the full dataset (Figure 2.1A). Using all data points, HSNE begins by constructing a neighborhood graph based on a user defined number of neighbors *k*. Next, HSNE defines a transition matrix T_0 based on two properties. First, the transition probability between two data points, i and j, is inversely proportional to the Euclidean distance between them. Second, each data point i is only allowed to transition to a data point j, if j belongs to the k-nearest-neighborhood of i, otherwise the transition probability is zero. The transition matrix encodes the intrascale similarities between data points.

To define the next scale S_1, HSNE selects representative data points from scale S_0, called landmarks. Landmarks on a given scale S_n represent a subset of (landmark) points at the previous scale S_{n-1}. To find the landmark point at scale S_1, HSNE performs many random walks of fixed length on the transition matrix T_0, starting from each data point at S_0. Next, HSNE records the number of random walks ending at each (landmark) point, reflecting the connectivity of each data point. Data points at S_0 with a connectivity above user defined threshold are selected as landmarks for S_1. As most data points do not meet this threshold, the new scale S_1 is more sparsely populated than the previous scale S_0 (Figure 2.1B).

To generate a new scale (say S_2) in the hierarchy, repeating the previous procedure cannot retain the original data structure. For instance, calculating another neighborhood graph on landmarks of scale S_1 will define neighbors with a short Euclidean distance that do not follow the original manifold (Figure 2.1B). To preserve the original data structure, HSNE uses a different concept, called the area of influence, to define neighborhoods for landmarks (Figure 2.1B-C). The area of influence of a landmark on scale S_n encodes the set of points, from the previous scale S_{n-1}, that can be represented by that landmark. Consequently, the area of influence matrix encodes the interscale similarities, where $A_n(i,j)$ is the probability that point i at scale S_{n-1} is well represented by landmark j at scale S_n. The similarities between the landmarks of scale S_n are calculated based on the overlap of their areas of influence on scale S_{n-1}, thus generating the transition matrix T_n for scale S_n (Figure 2.1D). As a result, each scale is sampled from the previous scale in such a way that the structure of the full data in the high dimensional space is retained.

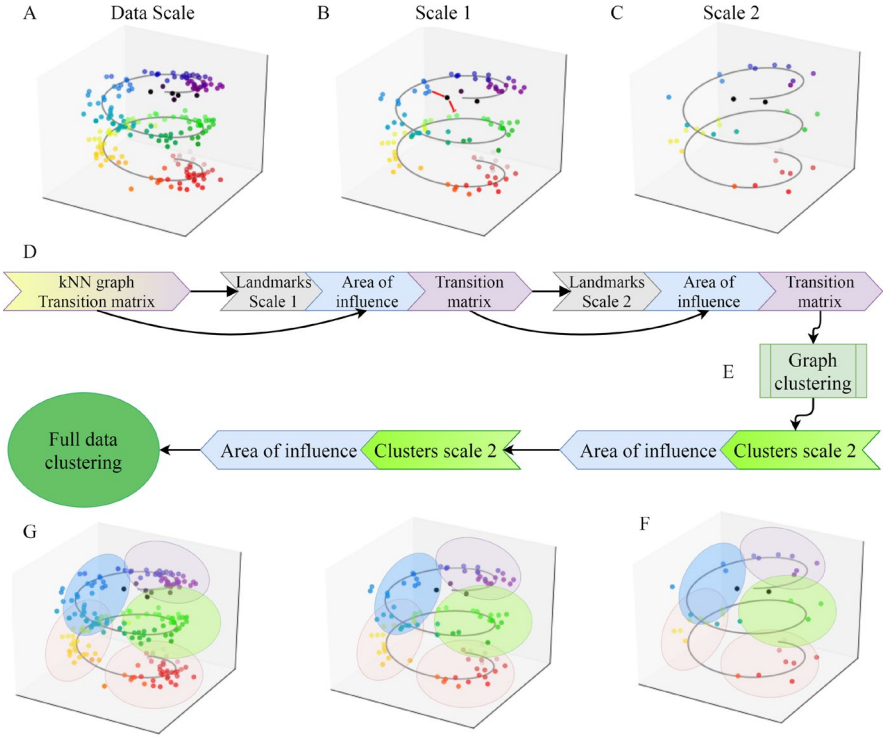


Figure 2.1 SCHNEL workflow. (A) In silico generated dataset: random data points on a spiral manifold in 3 dimensions. The data scale represents all data points within the dataset. The transition matrix is based on the kNN graph of all points. **(B)** At scale 1, highly connected points at the data scale are selected as landmarks. To keep the original manifold, the area of influence for each landmark is calculated, storing the impact/relationship of a landmark at scale 1 on/with the data points at the data scale. The red lines show that performing a kNN at scale 1 will find erroneous neighbors (in Euclidean space), i.e. neighbors that are more distant to each other than when following the spiral manifold. **(C)** Landmarks in scale 2 are subsequently a subset from the landmarks of scale 1. Each scale is sampled from the previous scale in such a way that the non-linear structure of the data in the high dimensional space is retained. **(D)** Flow of information in an HSNE hierarchy. Transition matrices are used to select landmarks for subsequent scales. At all scales (excluding the data scale), the transition matrix is calculated from the area of influence, which in turn is calculated based on the landmark selection process which is derived from the transition matrix at the previous scale. **(E)** Graph clustering is performed on a scale of choice, in this example scale 2. This is a computationally cheap operation since only a small subset of the data is clustered. **(F)** Cluster labels have been assigned to each landmark of scale 2, the labels are now propagated down the hierarchy to the data scale using the information encoded in the area of influence. **(G)** The full dataset now has cluster assignments, while only a fraction of the data was actually clustered.

2.2.1.2 Graph Clustering

At any scale of the hierarchy, the dataset can be clustered using a graph clustering to define the different clusters of points (cell populations for biological dataset) (Figure 2.1E). Depending on the number of (landmark) points on a scale this is feasible or not. In all our experiments all scales, except the data scale, were feasible, as the number of landmark points is at least an order of magnitude smaller than the number of data points in the full data set. We applied the Louvain Community detection, which is a heuristic algorithm that attempts to cluster the graph by maximizing the modularity[24]. Modularity is a graph property measuring how well clusters in a graph are separated[25]. Clustering is performed on the

transition matrices, and the results are propagated to the full dataset using the information encoded in the area of influence (Figure 2.1F-G), hence, for all scales, also a clustering of all data points is achieved.

2.2.1.3 Label Propagation

Once the landmarks of a given scale were clustered, these cluster labels are propagated down the hierarchy to label the full dataset. For this task, we used the area of influence (Supplementary Figure 2.1). The area of influence A_n at scale S_n is an i by j matrix, where j is the number of landmarks at scale S_n, and i is the number of landmarks/points at scale S_{n-1} preceding it. Each row is a probability distribution of point i at scale S_{n-1} being represented by landmarks at scale S_n.

We defined a cluster aggregated version of A_n named A_n^c, an i by c matrix, where c is the number of clusters obtained from clustering the j landmarks at scale S_n, and i is the number of landmarks/points at scale S_{n-1}. For each row i, the probabilities of landmarks (columns of A_n) belonging to the same cluster were summed. The inter-scale connection is defined as the maximum aggregate value of each row in A_n^c. The cluster label each row (data point that needs a label) received was the column (cluster) that had the highest aggregated probability in that row.

2.2.1.4 Implementation Details

We calculated the HSNE hierarchy and converted it to binary format using an adapted version of the High Dimensional Inspector C++ version 1.0.0 that saves the HSNE hierarchy to disk and omits the interactive tSNE. Settings for HSNE were: Beta threshold 1.5, number of neighbors 30, number of walks for landmark selection 200, number of scales round(log10(N/100)) where N is the number of points in the dataset.

The graph clustering is applied using the Python Louvain implementation version 0.6.1 (https://github.com/vtraag/louvain-igraph). The HSNE hierarchy is read using a custom written Python parser (https://github.com/paulderaadt/HSNE-clustering). The Louvain algorithm used the transition matrix values as weights, and modularityVertexPartition as maximization objective[26].

2.2.2 DATASETS

In this study, we applied and evaluated SCHNEL using nine different datasets: one popular machine learning benchmark dataset, seven publicly available cytometry datasets, and one single-cell RNA-sequencing dataset (Table 2.1).

The MNIST dataset contains handwritten digits that were scanned into a computer, each pixel has a value between 0 and 255 and is one of the 784 features of the dataset. It has ten different digits, the numbers 0-9 (http://yann.lecun.com/exdb/mnist/).

All the cytometry datasets are PBMC (Peripheral Blood Mononuclear Cells) or bone marrow samples measured with specific markers to analyze the immune system. The AML dataset is a small benchmark mass cytometry dataset, consisting of bone marrow samples from two healthy humans. The cells were manually gated by experts into 14 different cell populations[22]. The BMMC dataset is another small CyTOF dataset, containing a healthy human bone marrow sample from a single subject manually gated into 24 cell populations[22]. The Panorama dataset is a larger CyTOF dataset with 10 replicates of mouse bone marrow samples manually gated into 24 cell populations[23]. The HMIS dataset is an even larger CyTOF dataset, consisting of 47 human PBMC samples of healthy, Crohn's disease, and Celiac's

disease patients. There are no manually gated labels available. The HMIS dataset was analyzed and clustered using Cytosplore resulting into six major immune clusters[27]. The largest CyTOF dataset is Phenograph-Data, with more than 17 million cells derived from 21 human bone marrow healthy and leukemia individuals[22].

The Nilsson and Mossman datasets are both FC datasets from healthy humans. For both the Nilsson and Mossman datasets there are no full annotations available, only a very small (rare) population is annotated. For the Nilsson dataset, 358 (0.8%) cells are annotated as hematopoietic stem cells[28]. For the Mosmann dataset, 109 (0.03%) cells are labelled as CD4 memory T-cells[29].

Finally, we used the Mouse Nervous System (MNS) single-cell RNA-seq dataset, measuring the transcriptome wide expression of 19 different regions of the MNS clustered into 39 cell populations[30].

Prior to any analysis, The MNIST dataset and all the cytometry datasets were arcsinh transformed with a cofactor of 5, and all features/markers were used as input to SCHNEL. While the MNS dataset was first log-transformed, next we applied PCA retaining only the top 100 principle components, before inputting the data to SCHNEL.

Table 2.1 Description of the different datasets employed, showing: the total number of data points (cells or images), the number of features (pixels, proteins markers, or genes, for the MNIST dataset, cytometry dataset and scRNA-seq dataset, respectively), labels indicates the number of ground truth clusters of each dataset, and type of data.

Dataset	# of points	Features	Labels	Type
MNIST	60,000	784	10	Handwritten digits
AML	104,184	32	14	CyTOF
BMMC	81,747	12	24	CyTOF
Panorama	514,386	39	24	CyTOF
HMIS	3,553,596	28	6	CyTOF
Phenograph-Data	17.2 M	31	-	CyTOF
Nilsson	44,140	14	1	FC
Mosmann	396,460	13	1	FC
MNS	160,796	28,000	39	scRNA-seq

2.2.3 Evaluation Metrics

After propagating the cluster labels to all data points, the clustering can be evaluated using the full dataset. Although the task at hand is unsupervised clustering, we used three different supervised evaluation metrics, as for all datasets, except Phenograph-Data, we had manually annotated cell populations used as ground truth. The evaluation metrics are:

The adjusted Rand index (ARI), measuring the similarity between two sets of cluster label assignments[31]. It is adjusted for the chance of coincidentally correctly assigning a pair of data points to the same cluster. It lies in the range of [-1,1], where -1 is worse than random cluster assignment, and 1 is a perfect matching clustering.

The homogeneity score (HS), measuring the pureness of clusters, given a clustering result with a ground truth[32]. It is a score between [0,1], where 1 means that each cluster contains only data points of a single ground truth label.

The completeness score (CS), conversely measuring whether different ground truth groups are captured in distinct clusters[32]. It is also a score between [0,1], where 1 means that all members of a given ground truth label are assigned to the same cluster.

There is a trade-off between high homogeneity and high completeness: e.g. when several ground truth groups all get clustered into one cluster, completeness would be 1 and homogeneity would be 0. It is thus important to evaluate both measures simultaneously.

2.2.4 BENCHMARKING TOOLS

We benchmarked SCHNEL versus three popular clustering tools for cytometry data, Phenograph[22], X-shift[23], and FlowSOM[21]. Phenograph Version 1.5.2 was used with k=30 and default settings for all other parameters. (https://github.com/jacoblevine/PhenoGraph). X-shift was applied using number of neighbors = 20, Euclidean distance, noise threshold = 1, no normalization, no minimal Euclidean length, number of clusters K ranging from 225 to 15. The final number of clusters was determined with the built-in elbow method. Release 26-4-2018 was used (https://github.com/nolanlab/vortex/releases). FlowSOM was applied using x-dim=10, y-dim=10, compensate=False, transform=False, scale=False, maxMeta=40. FlowSOM version 1.1.4.1 was used, available as Bioconductor R package. All experiments were limited to run on a single core Intel Xeon X5670 2.93GHz CPU with 24 GB of memory (to be able to compare runtimes).

2.3 RESULTS

2.3.1 SCHNEL PRODUCES MEANINGFUL CLUSTERING

To evaluate the performance of SCHNEL, we first explored the MNIST dataset, as it has the advantage of easy interpretation of the resulting clusters (recall that the MNIST dataset consists of images of handwritten digits). With SCHNEL, we generated three hierarchical scales of the full data set, and provided a clustering for each scale. Clustering results as well as evaluation metrics are summarized in Table 2.2. For all scales, SCHNEL produced good clustering, with all evaluation metrics relatively high (> 0.8), despite the difference in the number of landmark points that were clustered at each scale, which vary by orders of magnitude. For instance, scale 3 had only 142 (landmark) data points (~0.24% of the full dataset) and SCHNEL was still able to produce good clustering with only one cluster less than the original MNIST dataset (9 out of 10).

Next, we applied SCHNEL to three CyTOF datasets AML, BMMC, and Panorama, and evaluated the clustering of each scale (Table 2.2). For the AML dataset, SCHNEL provided good clustering of scales S_1 and S_2, with the number of clusters close to the ground truth. While scales S_3 and S_4 showed under-clustering of the AML dataset, probably because there were very few landmark cells at these scales. For the BMMC dataset, SCHNEL under-clustered the data for all scales, with 10 clusters less than the ground truth for both scales S_1 and S_2, but still with relatively good performance. Also, we observed a similar clustering for both scales S_1 and S_2, with an order of magnitude difference in the number of cells between both scales (~17.04% and 2.45% of the full dataset, for S_1 and S_2, respectively). We obtained similar observations for the Panorama dataset. For scales S_1, S_2 and S_3 (~12.92% , 2.13% and 0.28% of the full dataset, respectively), SCHNEL produced good clustering, with the number of clusters close to the ground truth. S_4 showed under-clustering as it contained very few landmark cells (0.04% of the full dataset).

Table 2.2 Summary of SCHNEL results for MNIST, AML, BMMC and Panorama dataset, across all scales.

Dataset	Scale	# of points	# of clusters	ARI	HS	CS
MNIST	1	12,014	13	0.83	0.89	0.82
	2	1,759	11	0.87	0.90	0.87
	3	142	9	0.82	0.85	0.90
AML	1	16,031	16	0.72	0.93	0.79
	2	2,595	14	0.78	0.92	0.83
	3	292	10	0.80	0.92	0.85
	4	50	6	0.94	0.85	0.98
BMMC	1	13,932	14	0.92	0.88	0.94
	2	2,002	14	0.90	0.87	0.93
	3	118	9	0.79	0.79	0.96
Panorama	1	66,466	23	0.84	0.94	0.87
	2	10,943	21	0.84	0.94	0.86
	3	1,436	23	0.84	0.94	0.86
	4	217	11	0.91	0.85	0.93

The evaluation metrics give a general indication of the clustering quality, but do not show what factors drive the joining or splitting of manually annotated (ground truth) clusters. Therefore, for the interpretable MNIST dataset, we inspected the clustering at the most detailed scale (S_1) and the least detailed scale (S_3) (Supplementary Figure 2.2A-B). The contingency matrix at the detailed scale showed good cluster assignments for each digit, although ones and fives were split over multiple clusters (Supplementary Figure 2.2A). Further inspection of the average cluster images of the three clusters representing a handwritten 'one' (clusters 8, 10 and 11) revealed that their differences relate to the way a 'one' is written: straight written ones (Supplementary Figure 2.2C), ones written with a slant of 45 degrees clockwise (Supplementary Figure 2.2D), and ones written at angles in between (Supplementary Figure 2.2E). At the least detailed scale, the split clusters at S_1 were merged, but now also the images representing fours and nines were merged into a single cluster (Supplementary Figure 2.2B), due to an overlapping motif between them (Supplementary Figure 2.2F).

Next, we checked the contingency matrix of the AML dataset on the most detailed scale S_1 (Supplementary Figure 2.3A). The first seven clusters were homogeneous and represent some subsets of the major lineages. In cluster 7, SCHNEL merged CD16 positive and negative NK-cells, and in cluster 9 SCHNEL clumped many of the hematopoietic stem cells (HSPCs). We observed other instances where SCHNEL splits the ground truth classes into multiple clusters. Again, inspecting the cluster averages, which now can be represented as a heatmap of marker expressions, helps to reveal the reasons for splitting or merging ground truth clusters (Supplementary Figure 2.3B). For example, clusters 2 and 3 contained almost exclusively CD4 T-cells, which were distinct in their expression of CD7. Clusters 1 and 5 were most different in their expression of CD33. Clusters 4 and 6 were split on distinct expression of CD20. Additionally, cluster 14 contained CD4 and CD8 T-cells with very high expression of CD235ab. Although some of these clusters seem ambiguous, the overall results show the ability of SCHNEL to produce meaningful clusters using only a small fraction of the data (~15.39%).

For the Panorama dataset, SCHNEL produced almost identical clustering using less cells; at S_1 having 66,466 (12.92%) landmark cells, and at S_3 even with only 1,436 (0.28%) landmark

cells (Supplementary Figure 2.4). This illustrates the ability of SCHNEL to pick landmark cells that represent the dataset structure well.

2.3.2 SCHNEL OUTPERFORMS POPULAR CYTOMETRY CLUSTERING TOOLS

To further evaluated the performance of SCHNEL, we benchmarked SCHNEL against three popular clustering tools for cytometry data (FlowSOM, Phenograph and X-shift), using the three CyTOF datasets: AML, BMMC and Panorama. In terms of the ARI evaluation metric, SCHNEL outperformed other tools across all three datasets, except Phenograph for the BMMC dataset which performed similarly (Table 2.3). Further, SCHNEL showed better visual partitioning agreement compared to the ground truth manual annotations (Figure 2.2).

FlowSOM under-clustered the data using the default settings, in which case FlowSOM determines the optimal number of clusters automatically (Figure 2.2). However, FlowSOM was capable of a good clustering when the predefined number of clusters is close to the number of cell populations in the manual annotations (Supplementary Figure 2.5). But, generally, this information is not available beforehand. FlowSOM, on the other hand, was extremely fast (clustering the whole dataset under 10 minutes).

Phenograph showed similar partitioning to SCHNEL, but suffered from over-clustering in some cases providing very detailed small clusters (Figure 2.2). Speed-wise, SCHNEL was an order of magnitude faster than Phenograph across all cytometry datasets used in this study (Figure 2.3). For the 3.5 million HMIS dataset, Phenograph was even not able to complete the clustering after 7 days, at which point it was discontinued.

X-shift performed reasonably well on the AML and BMMC datasets, but found too many small clusters on the Panorama dataset. Generally, X-shift failed to define clear boundaries between clusters (Figure 2.2). X-shift was not timed because its implementation is a graphical user interface, but its computation time was around 30 minutes for the smaller datasets AML and BMMC, and up to 6 hours for the Panorama data. Similar to Phenograph, X-shift was not able to complete the clustering of the HMIS dataset after 7 days.

Table 2.3 Performance summary of SCHNEL versus FlowSOM, Phenograph and X-shift. Clusters indicates the number of clusters found for each combination of cluster tool and dataset, whereas ARI indicates the Adjusted Rand Index for that combination expressing how much it overlaps with the ground truth data.

		AML	BMMC	Panorama
SCHNEL	Clusters	14	14	21
	ARI	**0.78**	0.90	**0.84**
FlowSOM	Clusters	5	7	8
	ARI	0.68	0.62	0.44
Phenograph	Clusters	24	17	31
	ARI	0.61	**0.91**	0.67
X-shift	Clusters	21	19	70
	ARI	0.69	0.67	0.66

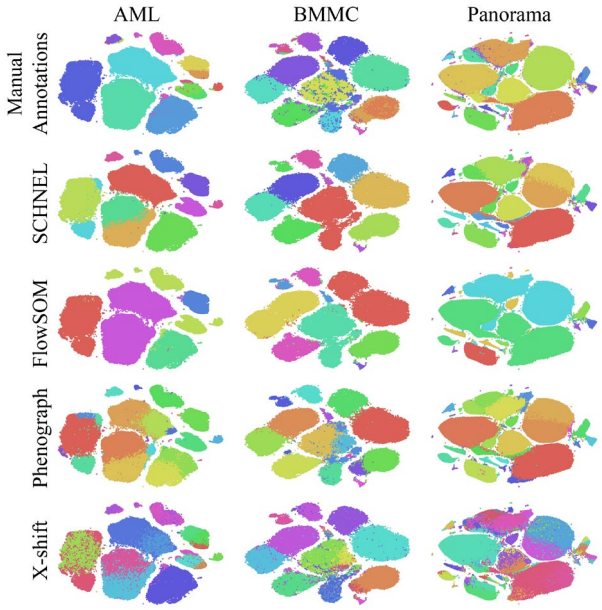


Figure 2.2 tSNE maps of AML, BMMC and Panorama datasets (columns) colored with different annotations (rows). Manual annotations indicate the ground truth labeling of the datasets. SCHNEL showed good visual agreement with the manual annotations. FlowSOM incorrectly merged different cell populations into mega clusters. Phenograph showed very detailed clustering. X-shift struggled to define clear cluster boundaries.

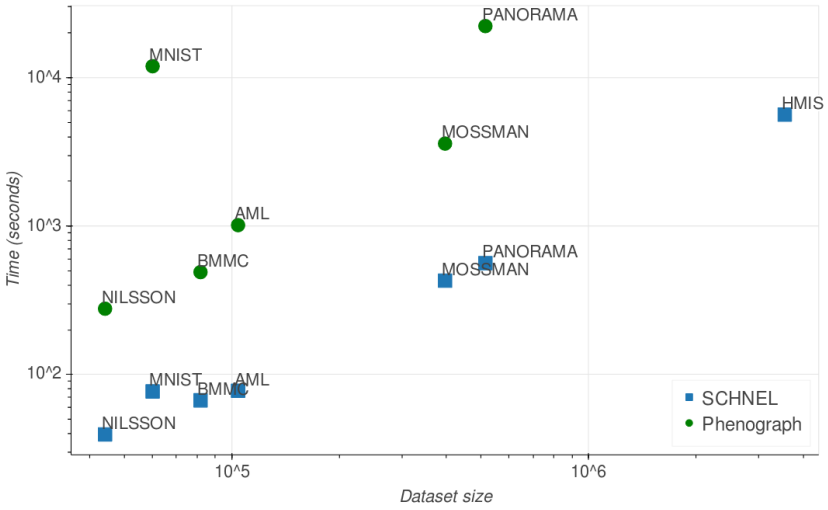


Figure 2.3 Computation time in seconds of SCHNEL and Phenograph with different dataset sizes. SCHNEL time is the clustering time of all scales in the hierarchy. Axes are log scaled.

2.3.3 SCHNEL SCALES TO LARGE DATASETS

SCHNEL was tested on datasets of different sizes to see how well it scales. Figure 2.4A shows the computation time of SCHNEL specified per task. Clustering the most detailed scale (S_1) was the most time-consuming operation. For the HMIS dataset this meant clustering 495,811 landmarks (similar as the entire Panorama dataset). Excluding this scale, the HMIS dataset could be clustered in roughly 50 minutes.

Further, we tested the scalability of SCHNEL to cluster the Phenograph-Data with 17.2 million cells, divided over 5 healthy and 16 leukemia individuals. In the original study, this dataset was analyzed per individual using the Phenograph clustering algorithm[22]. Using SCHNEL, we were able to pool all the cells from all individuals together and obtained a single clustering. We chose to represent the data at six different scales on top of the data scale. These scales contained 2.3M, 378K, 53K, 9K, 784 and 48 landmark cells, from the most detailed scale (S_1) to the least detailed scale (S_6), respectively. We skipped clustering S_1 as it is computationally very expensive. Clustering S_2 to S_6 resulted in 131, 133, 114, 47 and 5 clusters, respectively. Using the 47 cell clusters of S_5, we calculated the cluster frequencies across the 21 individuals (Figure 2.4B). Similar to the original study[22], we observed a homogeneous pattern across all healthy individuals, while the leukemia individuals had heterogeneous patterns. These results show the scalability of SCHNEL to cluster such large datasets and produce meaningful clustering.

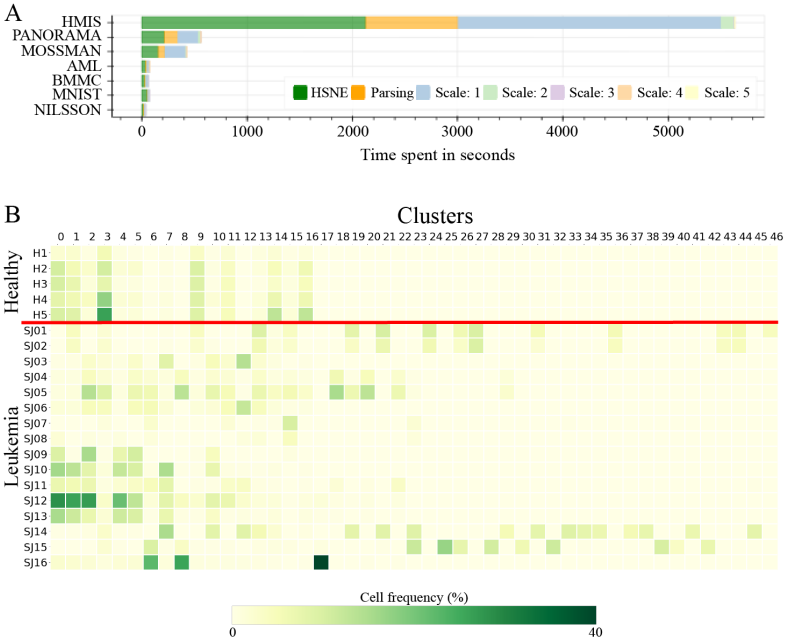


Figure 2.4 (A) Computation time of SCHNEL in seconds for different datasets. Different colors specify different steps in the SCHNEL algorithm. Green is calculating the HSNE hierarchies, orange is reading the HSNE hierarchy into Python, the other colors are times for clustering individual scales. Note that clustering scale 1 of the HMIS dataset (495,811 landmark cells) takes quite some time, showing the benefit of creating hierarchies and (only) clustering at higher scales having less landmarks. **(B)** Cluster frequencies across all the 21 individuals of the Phenograph-Data dataset. Clusters obtained from SCHNEL using scale 5. Red line separates between healthy and leukemia individuals.

2.3.4 SCHNEL DETECTS RARE CELL POPULATION

Different cell types are expected to have very different abundances and good clustering algorithms should be able to detect rare cell populations which are often interesting to study. We used the Mosmann and Nilsson datasets to test SCHNEL's sensitivity for detecting small populations. The Mosmann and Nilsson datasets both had manual annotations for only one rare cell population present within their full dataset. The Mosmann dataset contained a population of 109 memory CD4 T-cells. The Nilsson set contained 358 stem cells. Table 2.4 shows the sensitivity of SCHNEL for detecting these small populations. For both datasets, the cells belonging to the rare populations nicely clustered together at the various scales (CS), but for some scales these clusters also contained many other cells (Cluster size). For the Mosmann dataset, SCHNEL was able to capture the rare population in a single cluster without having many other cells at scales S_1 and S_3.

Table 2.4 Performance of SCHNEL for capturing rare populations in the Mosmann and Nilsson datasets. Cluster size indicates the size of the cluster in which most cells of the rare population were contained. Completeness Score (CS) indicates how many of the annotated rare cells in the original dataset were in the cluster containing most of the designated rare cells.

Dataset	Scale	# of cells	# of clusters	Cluster size	CS
Mosmann	1	77,787	23	181	0.94
	2	9,398	18	4,949	0.99
	3	1,090	14	173	0.93
	4	191	6	110,097	0.99
Nilsson	1	9,386	23	4,314	0.93
	2	1,354	17	3,269	0.93
	3	125	7	7,779	1

2.3.5 CLUSTERING SCRNA-SEQ DATA USING SCHNEL

After showing the potential of SCHNEL to cluster cytometry data, we tested the ability of SCHNEL to cluster scRNA-seq data which has many more features. We applied SCHNEL on the MNS dataset, using four scales on top of the data scale. Compared to the ground truth labels, the overall best clustering was obtained for scale 3 with 24 cell clusters, having an ARI of 0.68, HS of 0.83 and CS of 0.84. Similar to the MNIST dataset, we checked the clustering result details by calculating the contingency matrix, showing indeed a good agreement (Supplementary Figure 2.6), i.e. a clear one-to-one relation between SCHNEL clusters and ground truth labels. For example, cluster 10 with 'Microglia' and cluster 18 with 'Olfactory ensheathing cells'. In some cases, SCHNEL merged similar classes into one cluster. For instance, cluster 1 contained two 'Enteric' cells (glia and neurons) classes. Cluster 23 grouped three classes of 'Peripheral sensory neurons', while cluster 15 had two classes of 'Vascular cells'. Alternatively, SCHNEL did find two subtypes of 'Astrocytes' (cluster 3 and 7), and three subtypes of Oligodendrocytes (cluster 0,5 and 16).

2.4 DISCUSSION

SCHNEL provides a scalable fast solution for clustering large single cell data. Its novel approach utilizes HSNE for informed sampling of the data points using the concept of landmark selection and area of influence, which preserves the manifold structure of the full data. The sampling reduces the computational challenge of clustering many data points to a problem of clustering a subset of the data points that is at least an order of magnitude

smaller. The smaller subset can be quickly clustered by the Louvain algorithm, a graph-based clustering method. The manifold learning ensures that the sampled data points (landmark points) retain the same structure as the full dataset. As landmark points represent the data points, cluster labels can be easily propagated down the hierarchy to the full dataset.

Due to the informed sampling procedure, SCHNEL scales to large datasets. The results of the Phenograph-Data, HMIS and Panorama datasets showed that it is not necessary to cluster on the full dataset or even the most detailed scale in order to capture all clusters, even rare ones. In other words, a meaningful clustering can be obtained from a sampling of the data that is two, or more, orders of magnitude smaller than the full dataset. This gives SCHNEL the opportunity to cluster dataset sizes of up to millions of cells within workable timeframes.

When clustering the largest cytometry dataset, Phenograph-Data, SCHNEL was able to pool all cells from all individuals together in one clustering. Compared to a clustering of each individual separately (as done in the original Phenograph study), SCHNEL achieved two major advantages. Firstly, cell cluster frequencies across individuals can be directly applied as all clusters emerged from one clustering. This in contrast to clustering per individual which requires matching the clusters obtained across individuals to be able to compare their frequencies. Secondly, pooling all cells together helps to emphasize small rare cell populations, making them easier to detect. When analyzing per individual, rare cell population might be divided across individuals, resulting in too few cells to be detected as a separate population.

Using three CyTOF datasets, AML, BMMC, and Panorama, SCHNEL achieved similar or better performance compared to the tested existing tools. Moreover, SCHNEL did not require any pre-existing knowledge on how many clusters the data should contain. We observed an under-clustering of the BMMC dataset using SCHNEL, this may be due to the fact that the BMMC dataset contains 11 (out of 24) small cell populations with less than 1,000 cells. These small populations might not have enough representative landmarks in subsequent scales of the hierarchy.

When clustering the AML dataset, SCHNEL produced some interesting cell clusters that might have biological relevance. Clusters 2 and 3 separated the CD4 T-cells into two groups with different expression of CD7. CD7- CD4 T-cells have been reported before and can result from either ageing or prolonged immune system activation[33]. Clusters 4 and 6 were only distinct in their expression of CD20. Both clusters mainly contained mature B cells. This suggests that cluster 4 (CD20-) is a plasma B cell population, as CD20 is known to be highly expressed across all mature B-cells except plasma cells[34]. Finally, Cluster 14 contained a set of 62 CD4 T-cells and 43 CD8 T-cells. Normally these two proteins are mutually exclusive in mature T-cells. It could be possible that SCHNEL detected a rare subset of CD4+CD8+ T-cells. Therefore, formation of this cluster seems mostly driven by high expression of CD235ab, a red blood cell protein used to filter them out. It is suggested that these cells were not properly filtered out during manual gating.

Clustering the MNS scRNA-seq dataset showed that SCHNEL can handle large feature dimensions and produce meaningful clustering. These result shows that SCHNEL can be used as general clustering tool for single-cell data, not only cytometry data.

It is important to note that SCHNEL is a stochastic procedure, as the HSNE employs an approximated nearest neighbor search for generating the transition matrix on the data scale. This means that, although extremely similar, different hierarchies made from the same data with the same parameters will be slightly different. In addition, the Louvain clustering

algorithm is also stochastic because it chooses random nodes as candidates for merging when trying to optimize for modularity. Different runs of the Louvain method on the same graph may produce slightly different results.

The current implementation of SCHNEL provides clustering for all scales. This provides different level of details in the clustering, however, it limits the automation of the algorithm to produce one clustering of the data. Further improvements can automatically determine the scale providing the best clustering, which may also reduce the computation time.

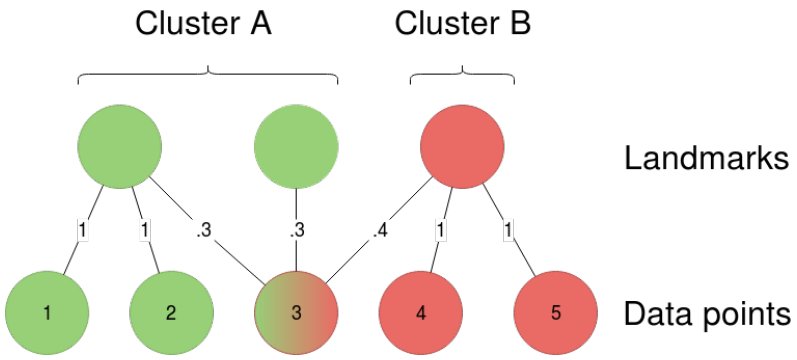
In conclusion, SCHNEL presents a reliable automated clustering tool for single-cell high-dimensional datasets. Using the HSNE, SCHNEL allows to perform graph clustering scalable to tens of millions of cells. Such clustering can be applied at different scales of the hierarchy, providing different level of detail.

BIBLIOGRAPHY

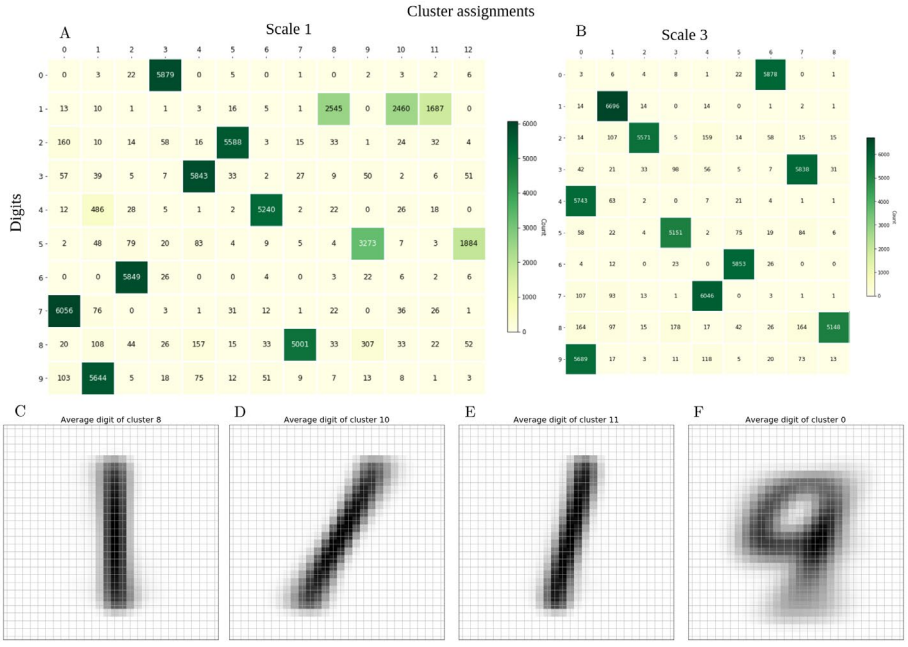
1. Picot, J., Guerin, C. L., Le Van Kim, C. & Boulanger, C. M. Flow cytometry: Retrospective, fundamentals and recent instrumentation. *Cytotechnology* **64,** 109–130 (2012).
2. Bandura, D. R. *et al.* Mass cytometry: Technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* **81,** 6813–6822 (2009).
3. Spitzer, M. H. & Nolan, G. P. Mass Cytometry: Single Cells, Many Features. *Cell* **165,** 780–791 (2016).
4. Virgo, P. F. & Gibbs, G. J. Flow cytometry in clinical pathology. *Ann. Clin. Biochem.* **49,** 17–28 (2012).
5. de Koning, C., Plantinga, M., Besseling, P., Boelens, J. J. & Nierkens, S. Immune Reconstitution after Allogeneic Hematopoietic Cell Transplantation in Children. *Biology of Blood and Marrow Transplantation* **22,** 195–206 (2016).
6. Stikvoort, A. *et al.* Combining flow and mass cytometry in the search for biomarkers in chronic graft-versus-host disease. *Front. Immunol.* **8,** (2017).
7. Hernandez-Martinez, J.-M., Vergara, E., Montes-Servín, E. & Arrieta, O. Interplay between immune cells in lung cancer: beyond T lymphocytes. *Transl. Lung Cancer Res.* **7,** S336–S340 (2018).
8. Aghaeepour, N. *et al.* Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods* **10,** 228–238 (2013).
9. Chester, C. & Maecker, H. T. Algorithmic Tools for Mining High-Dimensional Cytometry Data. *J. Immunol.* **195,** 773–779 (2015).
10. Weber, L. M. & Robinson, M. D. Comparison of Clustering Methods for High-Dimensional Single-Cell Flow and Mass Cytometry Data. *Cytom. A* **89,** 1084-1–96 (2016).
11. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn.* **9,** 2579–2605 (2008).
12. Shekhar, K., Brodin, P., Davis, M. M. & Chakraborty, A. K. Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE). *Proc. Natl. Acad. Sci. U. S. A.* **111,** 202–207 (2014).
13. Chen, H. *et al.* Cytofkit: A Bioconductor Package for an Integrated Mass Cytometry Data Analysis Pipeline. *PLoS Comput. Biol.* **12,** (2016).
14. Becher, B. *et al.* High-dimensional analysis of the murine myeloid cell system. *Nat. Immunol.* **15,** 1181–1189 (2014).
15. Maaten, L. van der. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15,** 3221–3245 (2014).
16. Pezzotti, N. *et al.* Approximated and User Steerable tSNE for Progressive Visual Analytics. *IEEE Trans. Vis. Comput. Graph.* **23,** 1739–1752 (2017).

17. Pezzotti, N. *et al.* GPGPU Linear Complexity t-SNE Optimization. *IEEE Trans. Vis. Comput. Graph.* **26,** 1172–1181 (2020).
18. Pezzotti, N., Höllt, T., Lelieveldt, B., Eisemann, E. & Vilanova, A. Hierarchical Stochastic Neighbor Embedding. in *Computer Graphics Forum (Proceedings of EuroVis 2016)* **35,** (2016).
19. Van Unen, V. *et al.* Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. *Nat. Commun.* **8,** 1–10 (2017).
20. Höllt, T. *et al.* Cytosplore : Interactive Immune Cell Phenotyping for Large Single-Cell Datasets. in *Computer Graphics Forum (Proceedings of EuroVis 2016)* **35,** (2016).
21. Van Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytom. Part A* **87,** 636–645 (2015).
22. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162,** 1–14 (2015).
23. Samusik, N., Good, Z., Spitzer, M. H., Davis, K. L. & Nolan, G. P. Automated mapping of phenotype space with single-cell data. *Nat. Methods* **13,** 493–496 (2016).
24. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008,** (2008).
25. Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 8577–8582 (2006).
26. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9,** (2019).
27. van Unen, V. *et al.* Mass Cytometry of the Human Mucosal Immune System Identifies Tissue- and Disease-Associated Immune Subsets. *Immunity* **44,** 1227–1239 (2016).
28. Rundberg Nilsson, A., Bryder, D. & Pronk, C. J. H. Frequency determination of rare populations by flow cytometry: A hematopoietic stem cell perspective. *Cytometry Part A* **83,** 721–727 (2013).
29. Mosmann, T. R. *et al.* SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, Part 2: Biological evaluation. *Cytom. Part A* **85,** 422–433 (2014).
30. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174,** 999–1014.e22 (2018).
31. Rand, W. M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66,** 846–850 (1971).
32. Rosenberg, A. & Hirschberg, J. V-Measure: A conditional entropy-based external cluster evaluation measure. in *EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* 410–420 (2007).
33. Reinhold, U. & Abken, H. CD4+CD7- T cells: A separate subpopulation of memory T cells? *J. Clin. Immunol.* **17,** 265–271 (1997).
34. Leandro, M. J. B-cell subpopulations in humans and their differential susceptibility to depletion with anti-CD20 monoclonal antibodies. *Arthritis Research and Therapy* **15,** (2013).

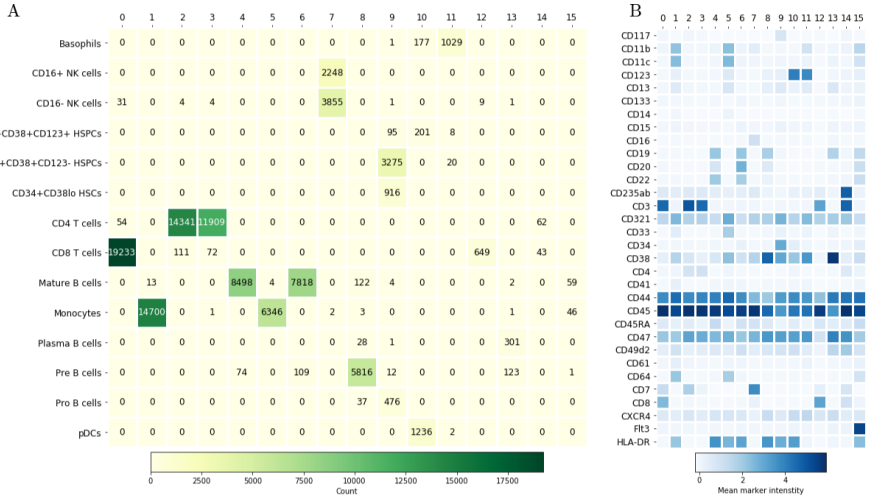
SUPPLEMENTARY MATERIALS



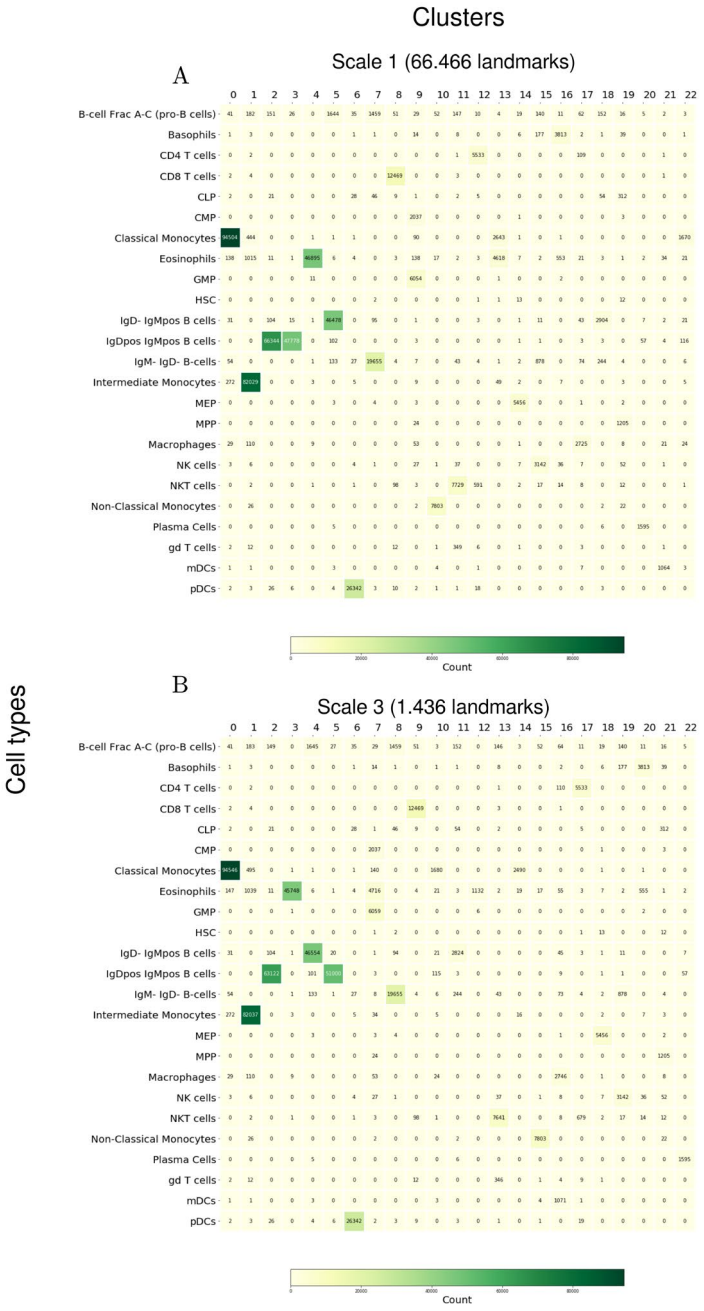
Supplementary Figure 2.1 Visual representation of propagating labels. Three landmarks are clustered into two groups. The area of influence is represented by the lines between the landmarks and the data points. The numbers on the edges represent the probability of a data point being well presented by a landmark. Propagating the landmark labels down to the data level is unambiguous for data point 1, 2, 4, and 5. Point 3 will be assigned to cluster A since it has a higher total probability (0.3 + 0.3 = 0.6 versus 0.4)



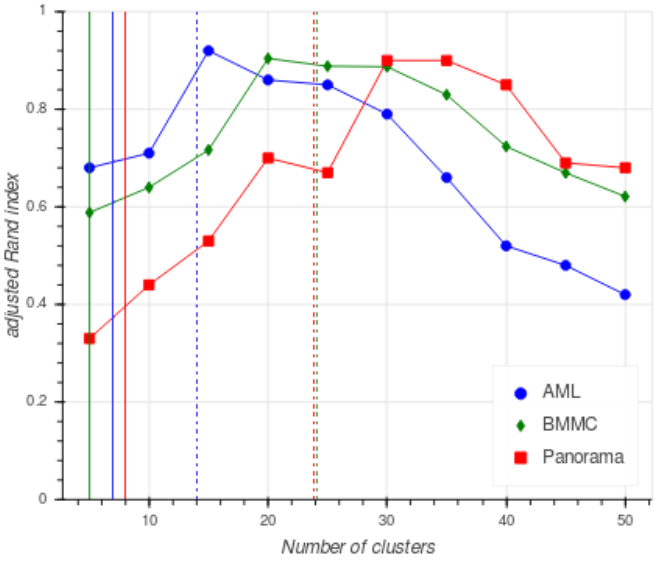
Supplementary Figure 2.2 (A) Contingency matrix of clustering scale 1 of the MNIST dataset. Ones were separated into three different clusters, and fives were split over two different clusters. **(B)** Contingency matrix of clustering scale 3 of the MNIST dataset. All digits were assigned their own cluster, except fours and nines which were merged. **(C-E)** The average pixel values for the clusters containing ones in scale 1, clusters 8, 10, and 11, respectively. **(F)** The average pixel values for cluster 0 on scale 3, which contained fours and nines.



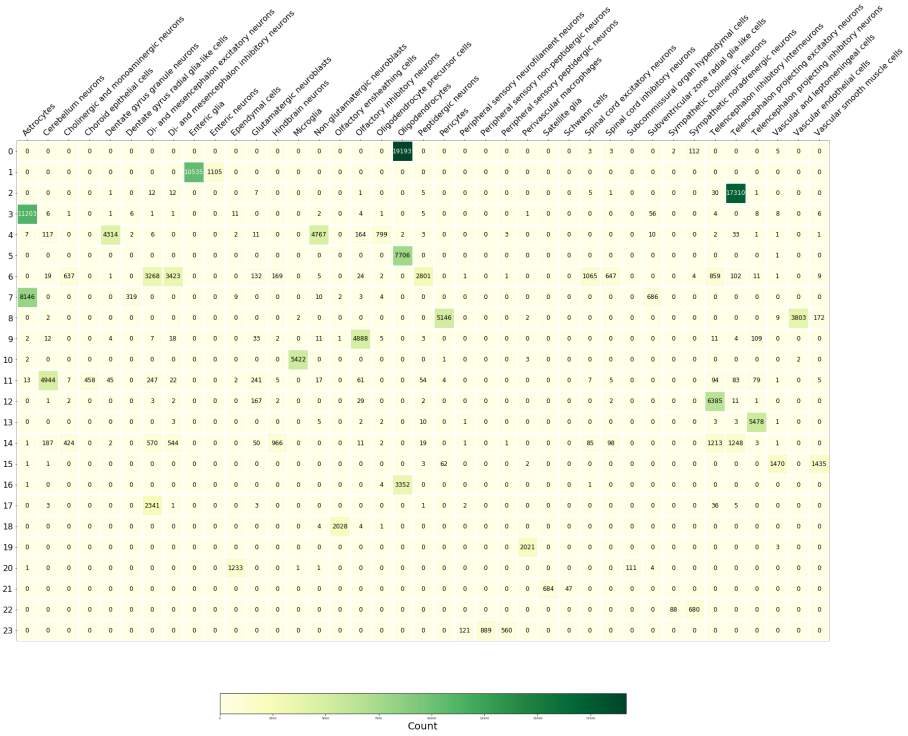
Supplementary Figure 2.3 (A) Contingency matrix of scale 1 compared to manual annotations of the AML dataset. **(B)** Heatmap showing the average marker expression per cluster for the results of AML scale 1. Darker color means higher expression of that protein for that cluster.



Supplementary Figure 2.4 Contingency matrices of the Panorama dataset clustering using SCHNEL on **(A)** scale 1 and **(B)** scale 3. Despite the very large difference in number of landmarks, the results are extremely similar.



Supplementary Figure 2.5 Adjusted Rand index of FlowSOM on AML, BMMC, and Panorama datasets. Ten runs were performed for each dataset with a different forced number of clusters, ranging from 5 to 50 with increments of 5. The solid vertical lines are the number of clusters FlowSOM outputs when it was allowed to optimize the number of clusters automatically. The dashed vertical lines indicate the number of cell populations per the manual annotations.



Supplementary Figure 2.6 Contingency matrix of the MNS dataset clustering using SCHNEL on scale 3.

CHAPTER 3

CYTOSPLORE-TRANSCRIPTOMICS: A SCALABLE INTERACTIVE FRAMEWORK FOR SINGLE-CELL RNA SEQUENCING DATA ANALYSIS

Tamim Abdelaal

Jeroen Eggermont

Thomas Höllt

Ahmed Mahfouz

Marcel J.T. Reinders

Boudewijn P.F. Lelieveldt

This chapter is published in: *Biorxiv* (2020), doi: 10.1101/2020.12.11.421883 (submitted).
Supplementary material is available online at:
https://www.biorxiv.org/content/10.1101/2020.12.11.421883v1.supplementary-material

The ever-increasing number of analyzed cells in Single-cell RNA sequencing (scRNA-seq) experiments imposes several challenges on the data analysis. Current analysis methods lack scalability to large datasets hampering interactive visual exploration of the data. We present Cytosplore-Transcriptomics, a framework to analyze scRNA-seq data, including data preprocessing, visualization and downstream analysis. At its core, it uses a hierarchical, manifold preserving representation of the data that allows the inspection and annotation of scRNA-seq data at different levels of detail. Consequently, Cytosplore-Transcriptomics provides interactive analysis of the data using low-dimensional visualizations that scales to millions of cells.

3.1 INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) is a valuable technology to identify the cellular composition of complex tissues[1]. Technological advances over the last decade resulted in a large increase in the acquired data size, scaling to millions of cells, raising major challenges for data analysis[2–4]. Current available tools, such as Seurat[5] and Scanpy[6], provide automated pipelines to analyze scRNA-seq datasets. Although these automated pipelines increase the reproducibility of analyses, they lack the possibility to interactively probe the data and intermediate results, which is essential since often the data analysis is largely exploratory.

Other tools offer interactive visualization and analysis for scRNA-seq data, including ASAP[7], cellxgene (https://github.com/chanzuckerberg/cellxgene), Granatum[8], Single Cell Explorer[9] and UCSC Cell Browser[10]. However, these tools do not scale to large datasets consisting of millions of cells. In addition, some tools are limited to a list of pre-loaded datasets, and do not allow users to explore, analyze and manually adjust annotations of their own data.

We present Cytosplore-Transcriptomics, a framework for interactive visual analysis and exploration of large scRNA-seq datasets consisting of millions of cells. Building on the principles of Cytosplore[11], we produce a hierarchical representation of the data using HSNE[12,13], which preserves the high-dimensional data manifold. We provide an interactive multiscale exploration of this hierarchy, starting from an abstract embedding containing fewer but representative cells for the global cellular composition, moving to more detailed embeddings of selections of cells on demand. The two-dimensional embeddings of the HSNE hierarchy can be used to cluster and define cell populations at different levels of the hierarchy, or to visualize the expression of selected genes and metadata across cells. Moreover, Cytosplore-Transcriptomics allows an interactive differential gene expression test between selected cell clusters.

3.2 METHODS

Cytosplore-Transcriptomics is able to perform data preprocessing, interactive data visualization, as well as downstream analysis such as clustering, cell type annotation and detecting differentially expressed genes across cell groups.

3.2.1 DATA INPUT AND FEATURE SELECTION

The user can provide data in various formats: (i) csv file containing genes as rows and cells as columns, (ii) hdf5 file including or excluding meta data, (iii) 10X sparse matrix format, or (iv) H5AD file containing a preprocessed Scanpy object. Additionally, meta-data can be uploaded separately in csv format. While uploading the data, CPM (count per million) normalization can be applied, as well as a log(x+1) or square root (sqrt) transformation.

Informative features/genes can be interactively selected to be used for the low-dimensional embedding (Figure 3.1A). First, the user may upload a list of genes to exclude from the analysis, such as mitochondrial genes. Next, highly variable genes can be selected by changing the selection threshold applied to the variance. In case of visualizing a previous analysis, it is also possible to upload a list of selected genes to be directly used for the embedding.

3.2.2 HIERARCHICAL VISUALIZATION

Once feature selection is performed, a hierarchical low-dimensional embedding of the data can be produced using HSNE. HSNE builds a hierarchy representing the dataset neighborhood in the high-dimensional feature space that preserves the manifold structure of the data, starting from the raw data points moving to multiple abstraction scales in a hierarchical way. The visualization of this hierarchy works in reverse order, by first showing a two-dimensional embedding of the highest scale in the hierarchy (overview scale) containing fewer, but representative, cells. Next, a more detailed embedding can be explored for a selected set of cells, by moving down through the hierarchy. In such a way, HSNE is scalable to millions of cells, without the need of downsampling, with the continuous possibility to explore the data hierarchy at more detailed scales. The number of scales is defined by the user and it is relative to the dataset size, it is recommended to set the number of scales to log10(N/100) where N is the total number of cells in the dataset. At any scale, gene expression and metadata can be overlaid on the low-dimensional embedding.

3.2.3 CLUSTERING AND ANNOTATION

To define different cell populations in the data, Cytosplore-Transcriptomics provides two different clustering methods, density-based and graph-based clustering. The density-based clustering relies on the low-dimensional embedding, where the layout of the cells indicates the similarity in the high-dimensional feature space. Based on the density representation of the embedding, unsupervised Gaussian Mean Shift (GMS) clustering can be applied to define different cell clusters. On the other hand, graph-based SCHNEL clustering[14] can be applied independently from the low-dimensional embedding, as the SCHNEL clustering applies the (Louvain or Leiden) community detection algorithm[15,16] on the neighborhood graph of each scale in the hierarchy. Moreover, Cytosplore-Transcriptomics allows the user to manually select and annotate (or correct annotations of) a set of cells of interest.

3.2.4 DIFFERENTIAL GENE EXPRESSION

Cytosplore-Transcriptomics provides an interactive differential expression test between different groups of cells (Wilcoxon rank-sum test with Bonferroni multiple testing correction). These groups can be cell clusters or a set of manually selected cells. The set of differentially expressed genes (DEgenes) can be provided either between two groups of cells, or one group versus the remaining cells. Next, the user may pick any of the DEgenes to visualize its expression level on the current embedding.

3.3 CASE STUDY

To illustrate the features of Cytosplore-Transcriptomics, we chose the mouse whole cortex and hippocampus dataset from the Allen Institute (https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-whole-cortex-and-hippocampus-10x), representing a relatively large scRNA-seq dataset with over a million cells having diverse cellular populations. We downloaded the original data files, and converted it to one hdf5 file including the metadata

(https://doi.org/10.5281/zenodo.4317397). We used Cytosplore-Transcriptomics for visual exploration of this data. First, the data with corresponding metadata is loaded, and we applied CPM normalization and a log(x+1) transformation to the data. Next, we excluded mitochondrial and sex related genes, and selected the top 3,078 highly variable genes for further analysis. An HSNE hierarchy with four scales (data scale + 3 higher scales) was computed and an overview embedding (scale 3) showing only 1,970 cells (0.18% of the full dataset) was visualized (Figure 3.1B). This scale shows the overall structure of the dataset with a clear separation between 34 cell populations identified in the metadata. To reveal more detailed structures (Figure 3.1C), we zoomed one scale deeper into the hierarchy, examining the embedding of scale 2 comprising 11,417 cells (1.04% of the full dataset). An interesting feature of the hierarchical exploration is the ability to zoom into a specific set of cells. For instance, in Figure 3.1D, we focused on a small group of cells from the hippocampus (highlighted in red in Figure 3.1B) and generated a separate, more detailed embedding of these cells. This new embedding clearly reveals the heterogeneity in the cellular composition of this specific group of cells, as several smaller subpopulations can be identified from different hippocampal regions, including CA1, retrohippocampal and prosubiculum. Next, we applied the SCHNEL clustering to the 1,970 cells at scale 3, producing 20 cell clusters (Supplementary Figure 3.1A). We quantified the agreement of this clustering result with the 34 labels from the metadata using the adjusted Rand index (ARI), measuring the similarity between two different groupings of cells, and obtained an ARI of 0.75 (1 being perfectly similar). We found 10 more clusters when applying SCHNEL to the more detailed scale 2 (Supplementary Figure 3.1A), with a total of 30 cell clusters that collectively have an ARI of 0.72 compared to the metadata labels. Finally, we calculated the DEgenes between two adjacent cell populations, Vip and Sncg neurons (highlighted in blue and green, respectively, in Figure 3.1C), to reveal the driving genes for these cellular populations. We zoomed into these two populations generating a separate embedding (Figure 3.1E), and overlaid the expression of the top DEgenes for each population, showing that *Caln1* is differentially expressed in the Vip neurons, while *Cck* is differentially expressed in the Sncg neurons (Supplementary Figure 3.1B). In total, 2,845 DEgenes are obtained (corrected p-value < 0.05, absolute average log2 fold-change > 1), each with their relevant statistics, including mean expression in each population, mean difference between populations, original and corrected p-values (Figure 3.1F).

3.4 CONCLUSIONS

We developed Cytosplore-Transcriptomics, a standalone tool that facilitates interactive visual exploration and analysis of large scRNA-seq datasets consisting of millions of cells, while preserving the manifold structure of the full data. In addition, it offers many interactive features including, feature selection, clustering and differential gene expression.

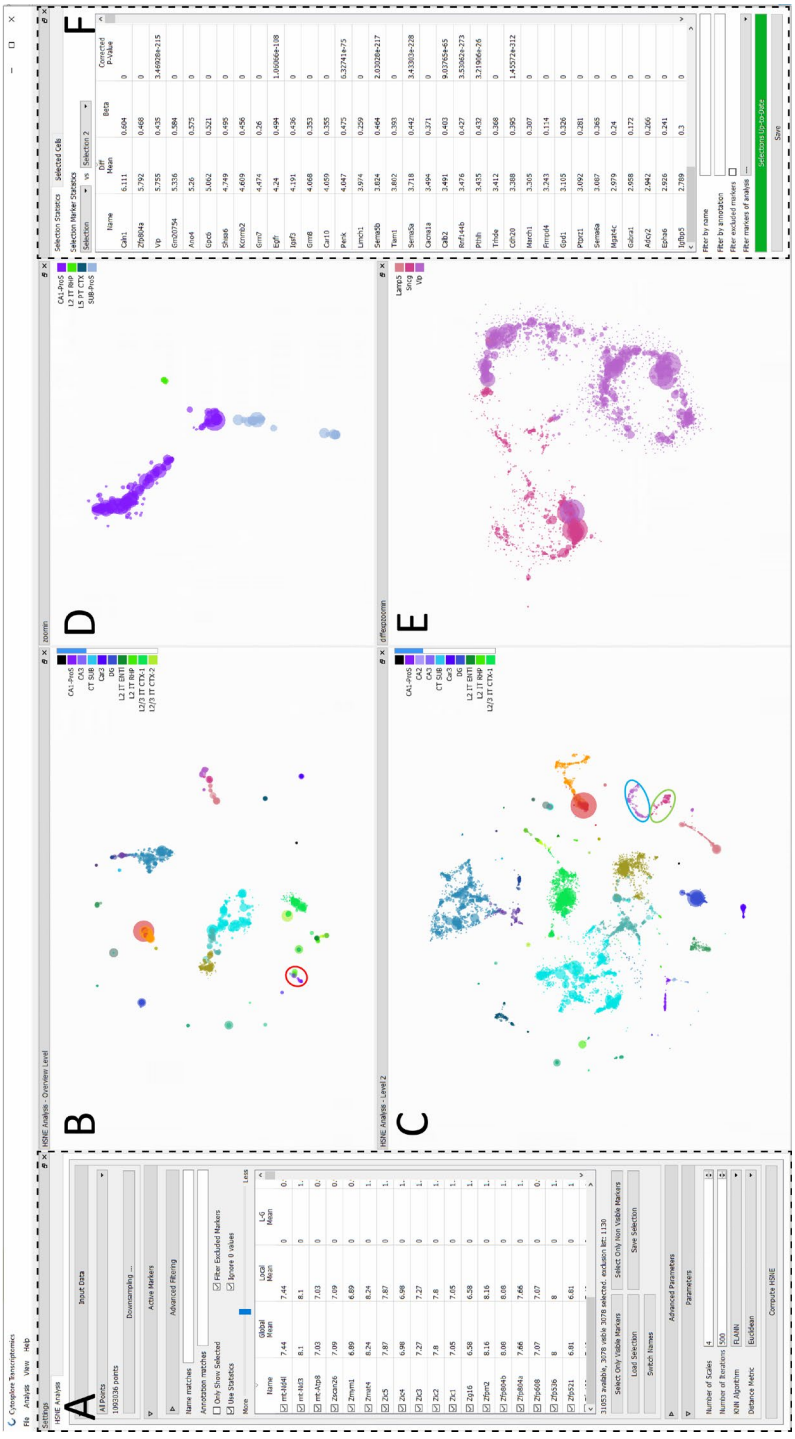


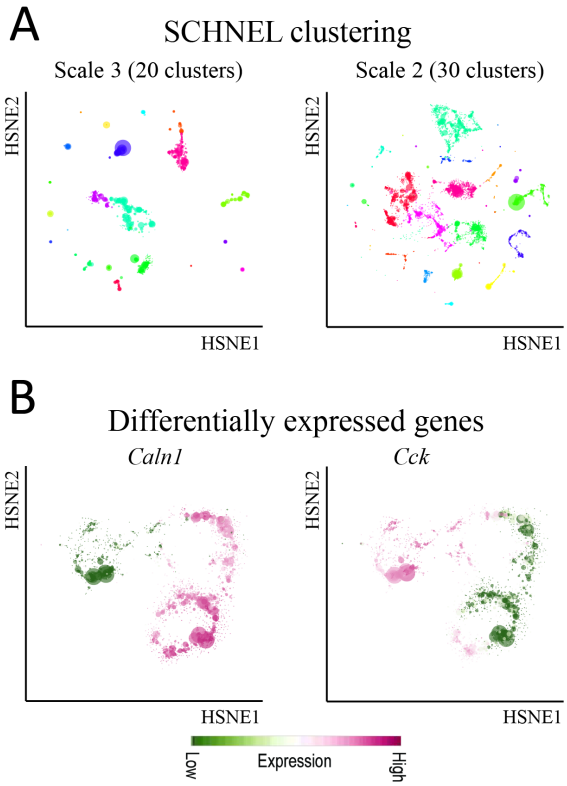
Figure 3.1 Cytosplore-Transcriptomics software. (A) HSNE analysis settings panel, where feature selection can be performed, and HSNE parameters can be selected. **(B)** Exploration of the data hierarchy by first showing the HSNE embedding of the overview scale with only 1,970 cells (0.18% of the total

number of cells). **(C)** Zooming one scale deeper into the hierarchy to scale 2 having 11,417 cells (1.04% of the total number of cells). **(D)** HSNE embedding zooming into a specific group of hippocampus cells, highlighted in red in (B), further revealing the cellular diversity within this group. **(E)** HSNE embedding zooming on the Vip and Sncg neurons, used for differential expression analysis, highlighted in blue and green, respectively, in (C)**.** All plots are colored according to the labels from the metadata. **(F)** Differential expression panel showing all genes with their corresponding statistics.

BIBLIOGRAPHY

1. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566,** 496–502 (2019).
2. Angerer, P. *et al.* Single cells make big data: New challenges and opportunities in transcriptomics. *Current Opinion in Systems Biology* **4,** 85–91 (2017).
3. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols* **13,** 599–604 (2018).
4. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biology* **21,** (2020).
5. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888–1902.e21 (2019).
6. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19,** (2018).
7. Gardeux, V., David, F. P. A., Shajkofci, A., Schwalie, P. C. & Deplancke, B. ASAP: A web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. *Bioinformatics* **33,** 3123–3125 (2017).
8. Zhu, X. *et al.* Granatum: A graphical single-cell RNA-Seq analysis pipeline for genomics scientists. *Genome Med.* **9,** 1–12 (2017).
9. Feng, D., Whitehurst, C. E., Shan, D., Hill, J. D. & Yue, Y. G. Single Cell Explorer, collaboration-driven tools to leverage large-scale single cell RNA-seq data. *BMC Genomics* **20,** 1–8 (2019).
10. Speir, M. L. *et al.* UCSC Cell Browser: Visualize Your Single-Cell Data. *bioRxiv* **2,** (2020).
11. Höllt, T. *et al.* Cytosplore : Interactive Immune Cell Phenotyping for Large Single-Cell Datasets. in *Computer Graphics Forum (Proceedings of EuroVis 2016)* **35,** (2016).
12. Pezzotti, N., Höllt, T., Lelieveldt, B., Eisemann, E. & Vilanova, A. Hierarchical Stochastic Neighbor Embedding. in *Computer Graphics Forum (Proceedings of EuroVis 2016)* **35,** (2016).
13. Van Unen, V. *et al.* Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. *Nat. Commun.* **8,** 1–10 (2017).
14. Abdelaal, T., de Raadt, P., Lelieveldt, B. P. F., Reinders, M. J. T. & Mahfouz, A. SCHNEL: scalable clustering of high dimensional single-cell data. *Bioinformatics* **36,** i849–i856 (2020).
15. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008,** (2008).
16. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9,** (2019).

SUPPLEMENTARY MATERIALS



Supplementary Figure 3.1 (A) SCHNEL clustering of scale 3 (left plot) and scale 2 (right plot) producing 20 and 30 cell clusters, respectively. Colors represent different cell clusters. **(B)** Expression profiles of top DEgenes between two adjacent cell populations, Vip and Sncg neurons, highlighted in blue and green, respectively, in Figure 3.1C. The gene expression is overlaid on the HSNE embedding zooming on the two populations of interest (Figure 3.1E). *Caln1* is differentially expressed in the Vip neurons, while *Cck* is differentially expressed in the Sncg neurons.

CHAPTER 4

A COMPARISON OF AUTOMATIC CELL IDENTIFICATION METHODS FOR SINGLE-CELL RNA SEQUENCING DATA

Tamim Abdelaal*

Lieke Michielsen*

Davy Cats

Dylan Hoogduin

Hailing Mei

Marcel J.T. Reinders

Ahmed Mahfouz

This chapter is published in: *Genome Biology* (2019) 20: 194, doi: 10.1186/s13059-019-1795-z.
Supplementary material is available online at:
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1795-z#Sec36
*Equal contribution

Single-cell transcriptomics is rapidly advancing our understanding of the cellular composition of complex tissues and organisms. A major limitation in most analysis pipelines is the reliance on manual annotations to determine cell identities, which are time-consuming and irreproducible. The exponential growth in the number of cells and samples has prompted the adaptation and development of supervised classification methods for automatic cell identification.

Here, we benchmarked 22 classification methods that automatically assign cell identities including single-cell-specific and general-purpose classifiers. The performance of the methods is evaluated using 27 publicly available single-cell RNA sequencing datasets of different sizes, technologies, species, and levels of complexity. We use 2 experimental setups to evaluate the performance of each method for within dataset predictions (intra-dataset) and across datasets (inter-dataset) based on accuracy, percentage of unclassified cells, and computation time. We further evaluate the methods' sensitivity to the input features, number of cells per population, and their performance across different annotation levels and datasets. We find that most classifiers perform well on a variety of datasets with decreased accuracy for complex datasets with overlapping classes or deep annotations. The general-purpose support vector machine classifier has overall the best performance across the different experiments.

In conclusion, we present a comprehensive evaluation of automatic cell identification methods for single-cell RNA sequencing data. All the code used for the evaluation is available on GitHub (https://github.com/tabdelaal/scRNAseq_Benchmark). Additionally, we provide a Snakemake workflow to facilitate the benchmarking and to support the extension of new methods and new datasets.

4.1 BACKGROUND

Single-cell RNA sequencing (scRNA-seq) provides unprecedented opportunities to identify and characterize the cellular composition of complex tissues. Rapid and continuous technological advances over the past decade have allowed scRNA-seq technologies to scale to thousands of cells per experiment[1]. A common analysis step in analyzing single-cell data involves the identification of cell populations presented in a given dataset. This task is typically solved by unsupervised clustering of cells into groups based on the similarity of their gene expression profiles, followed by cell population annotation by assigning labels to each cluster. This approach proved very valuable in identifying novel cell populations and resulted in cellular maps of entire cell lineages, organs, and even whole organisms[2–7]. However, the annotation step is cumbersome and time-consuming as it involves manual inspection of cluster-specific marker genes. Additionally, manual annotations, which are often not based on standardized ontologies of cell labels, are not reproducible across different experiments within and across research groups. These caveats become even more pronounced as the number of cells and samples increases, preventing fast and reproducible annotations.

To overcome these challenges, a growing number of classification approaches are being adapted to automatically label cells in scRNA-seq experiments. scRNA-seq classification methods predict the identity of each cell by learning these identities from annotated training data (e.g., a reference atlas). scRNA-seq classification methods are relatively new compared to the plethora of methods addressing different computational aspects of single-cell analysis (such as normalization, clustering, and trajectory inference). However, the number of classification methods is rapidly growing to address the aforementioned challenges[8,9]. While all scRNA-seq classification methods share a common goal, i.e., accurate annotation of cells,

they differ in terms of their underlying algorithms and the incorporation of prior knowledge (e.g., cell type marker gene tables).

In contrast to the extensive evaluations of clustering, differential expression, and trajectory inference methods[10–12], there is currently one single attempt comparing methods to assign cell type labels to cell clusters[13]. The lack of a comprehensive comparison of scRNA-seq classification methods leaves users without indications as to which classification method best fits their problem. More importantly, a proper assessment of the existing approaches in comparison with the baseline methods can greatly benefit new developments in the field and prevent unnecessary complexity.

Here, we benchmarked 22 classification methods to automatically assign cell identities including single-cell-specific and general-purpose classifiers. The methods were evaluated using 27 publicly available single-cell RNA sequencing datasets of different sizes, technologies, species, and complexity. The performance of the methods was evaluated based on their accuracy, percentage of unclassified cells, and computation time. We performed several experiments to cover different levels of challenge in the classification task and to test specific features or tasks such as the feature selection, scalability, and rejection experiments. We evaluated the classification performance through two experimental setups: (1) intra-dataset in which we applied 5-fold cross-validation within each dataset and (2) inter-dataset involving across datasets comparisons. The inter-dataset comparison is more realistic and more practical, where a reference dataset (e.g., atlas) is used to train a classifier which can then be applied to identify cells in new unannotated datasets. However, in order to perform well across datasets, the classifier should also perform well using the intra-dataset setup on the reference dataset. The intra-dataset experiments, albeit artificial, provide an ideal scenario to evaluate different aspects of the classification process (e.g., feature selection, scalability, and different annotation levels), regardless of the technical and biological variations across datasets. In general, most classifiers perform well across all datasets in both experimental setups (inter- and intra-dataset), including the general-purpose classifiers. In our experiments, incorporating prior knowledge in the form of marker genes does not improve the performance. We observed large variation across different methods in the computation time and classification performance in response to changing the input features and the number of cells. Our results highlight the general-purpose support vector machine (*SVM*) classifier as the best performer overall.

4.2 RESULTS

4.2.1 BENCHMARKING AUTOMATIC CELL IDENTIFICATION METHODS (INTRA-DATASET EVALUATION)

We benchmarked the performance and computation time of all 22 classifiers (Table 4.1) across 11 datasets used for intra-dataset evaluation (Table 4.2). Classifiers were divided into two categories: (1) supervised methods which require a training dataset labeled with the corresponding cell populations in order to train the classifier or (2) prior-knowledge methods, for which either a marker gene file is required as an input or a pretrained classifier for specific cell populations is provided.

Table 4.1 Automatic cell identification methods included in this study

Name	Version	Language	Underlying classifier	Prior knowledge	Rejection option	Ref.
Garnett	0.1.4	R	Generalized linear model	Yes	Yes	[14]
Moana	0.1.1	python	SVM with linear kernel	Yes	No	[15]
DigitalCellSorter	Github version: e369a34	python	Voting based on cell type markers	Yes	No	[16]
SCINA	1.1.0	R	Bimodal distribution fitting for marker-genes	Yes	No	[17]
scVI	0.3.0	python	Neural Network	No	No	[18]
Cell-BLAST	0.1.2	python	Cell-to-cell similarity	No	Yes	[19]
ACTINN	GitHub version: 563bcc1	python	Neural Network	No	No	[20]
LAmbDA	GitHub version: 3891d72	python	Random Forest	No	No	[21]
Scmap-cluster	1.5.1	R	Nearest median classifier	No	Yes	[22]
Scmap-cell	1.5.1	R	kNN	No	Yes	[22]
scPred	0.0.0.9000	R	SVM with radial kernel	No	Yes	[23]
CHETAH	0.99.5	R	Correlation to training set	No	Yes	[24]
CaSTLe	Github version: 258b278	R	Random Forest	No	No	[25]
SingleR	0.2.2	R	Correlation to training set	No	No	[26]
scID	0.0.0.9000	R	LDA	No	Yes	[27]
singleCellNet	0.1.0	R	Random Forest	No	No	[28]
LDA	0.19.2	python	LDA	No	No	[29]
NMC	0.19.2	python	NMC	No	No	[29]

RF	0.19.2	python	RF (50 trees)	No	No	[29]
SVM	0.19.2	python	SVM (linear kernel)	No	No	[29]
$SVM_{rejection}$	0.19.2	python	SVM (linear kernel)	No	Yes	[29]
kNN	0.19.2	python	kNN (k = 9)	No	No	[29]

Table 4.2 Overview of the datasets used during this study

Dataset	No. of cells	No. of genes	No. of cell populations (>10 cells)	Description	Protocol	Ref.
Baron (Mouse)[a]	1,886	14,861	13 (9)	Mouse Pancreas	inDrop	[30]
Baron (Human)[a,b]	8,569	17,499	14 (13)	Human Pancreas	inDrop	[30]
Muraro[a,b]	2,122	18,915	9 (8)	Human Pancreas	CEL-Seq2	[31]
Segerstolpe[a,b]	2,133	22,757	13 (9)	Human Pancreas	SMART-Seq2	[32]
Xin[a,b]	1,449	33,889	4 (4)	Human Pancreas	SMARTer	[33]
CellBench 10X[a,b]	3,803	11,778	5 (5)	Mixture of five human lung cancer cell lines	10X Chromium	[34]
CellBench CEL-Seq2[a,b]	570	12,627	5 (5)	Mixture of five human lung cancer cell lines	CEL-Seq2	[34]
TM[a]	54,865	19,791	55 (55)	Whole Mus musculus	SMART-Seq2	[6]
AMB[a]	12,832	42,625	4/22/110 (3/16/92)	Primary mouse visual cortex	SMART-Seq v4	[35]
Zheng sorted[a]	20,000	21,952	10 (10)	FACS sorted PBMC	10X Chromium	[36]
Zheng 68K[a]	65,943	20,387	11 (11)	PBMC	10X Chromium	[36]
VISp[b] (Mouse)	12,832	42,625	3/36 (3/34)	Primary Visual Cortex	SMART-Seq v4	[35]
ALM[b] (Mouse)	8,758	42,461	3/37 (3/34)	Anterior Lateral Motor Area	SMART-Seq v4	[35]
MTG[b] (Human)	14,636	16,161	3/35 (3/34)	Middle Temporal Gyrus	SMART-Seq v4	[37]

PbmcBench pbmc1.10Xv2[b]	6,444	33,694	9 (9)	PBMC	10X version 2	[38]
PbmcBench pbmc1.10Xv3[b]	3,222	33,694	8 (8)	PBMC	10X version 3	[38]
PbmcBench pbmc1.CL[b]	253	33,694	7 (7)	PBMC	CEL-Seq2	[38]
PbmcBench pbmc1.DR[b]	3,222	33,694	9 (9)	PBMC	Drop-Seq	[38]
PbmcBench pbmc1.iD[b]	3,222	33,694	7 (7)	PBMC	inDrop	[38]
PbmcBench pbmc1.SM2[b]	253	33,694	6 (6)	PBMC	SMART-Seq2	[38]
PbmcBench pbmc1.SW[b]	3,176	33,694	7 (7)	PBMC	Seq-Well	[38]
PbmcBench pbmc2.10Xv[,b]	3,362	33,694	9 (9)	PBMC	10X version 2	[38]
PbmcBench pbmc2.CL[b]	273	33,694	5 (5)	PBMC	CEL-Seq2	[38]
PbmcBench pbmc2.DR[b]	3,362	33,694	6 (6)	PBMC	Drop-Seq	[38]
PbmcBench pbmc2.iD[b]	3,362	33,694	9 (9)	PBMC	inDrop	[38]
PbmcBench pbmc2.SM2[b]	273	33,694	6 (6)	PBMC	SMART-Seq2	[38]
PbmcBench pbmc2.SW[b]	551	33,694	4 (4)	PBMC	Seq-Well	[38]

a: used for intra-dataset evaluation
b: used for inter-dataset evaluation

The datasets used in this study vary in the number of cells, genes, and cell populations (annotation level), in order to represent different levels of challenges in the classification task and to evaluate how each classifier performs in each case (Table 4.2). They include relatively typical sized scRNA-seq datasets (1500–8500 cells), such as the 5 pancreatic datasets (Baron Mouse, Baron Human, Muraro, Segerstolpe, and Xin), which include both mouse and human pancreatic cells and vary in the sequencing protocol used. The Allen Mouse Brain (AMB) dataset is used to evaluate how the classification performance changes when dealing with different levels of cell population annotation as the AMB dataset contains three levels of annotations for each cell (3, 16, or 92 cell populations), denoted as AMB3, AMB16, and AMB92, respectively. The Tabula Muris (TM) and Zheng 68K datasets represent relatively large scRNA-seq datasets (> 50,000 cells) and are used to assess how well the classifiers scale with large datasets. For all previous datasets, cell populations were obtained through clustering. To assess how the classifiers perform when dealing with sorted populations, we included the CellBench dataset and the Zheng sorted dataset, representing sorted populations for lung cancer cell lines and peripheral blood mononuclear cells (PBMC), respectively. Including the Zheng sorted and Zheng 68K datasets allows the benchmarking of 4 prior-knowledge classifiers, since the marker gene files or pretrained classifiers are available for the 4 classifiers for PBMCs.

4.2.2 ALL CLASSIFIERS PERFORM WELL IN INTRA-DATASET EXPERIMENTS

Generally, all classifiers perform well in the intra-dataset experiments, including the general-purpose classifiers (Figure 4.1). However, *Cell-BLAST* performs poorly for the Baron Mouse and Segerstople pancreatic datasets. Further, *scVI* has low performance on the deeply annotated datasets TM (55 cell populations) and AMB92 (92 cell populations), and *kNN* produces low performance for the Xin and AMB92 datasets.

For the pancreatic datasets, the best-performing classifiers are *SVM*, $SVM_{rejection}$, *scPred*, *scmapcell*, *scmapcluster*, *scVI*, *ACTINN*, *singleCellNet*, *LDA*, and *NMC*. *SVM* is the only classifier to be in the top five list for all five pancreatic datasets, while *NMC*, for example, appears only in the top five list for the Xin dataset. The Xin dataset contains only four pancreatic cell types (alpha, beta, delta, and gamma) making the classification task relatively easy for all classifiers, including *NMC*. Considering the median F1-score alone to judge the classification performance can be misleading since some classifiers incorporate a rejection option (e.g., $SVM_{rejection}$, *scmapcell*, *scPred*), by which a cell is assigned as "unlabeled" if the classifier is not confident enough. For example, for the Baron Human dataset, the median F1-score for $SVM_{rejection}$, *scmapcell*, *scPred*, and *SVM* is 0.991, 0.984, 0.981, and 0.980, respectively (Figure 4.1A). However, $SVM_{rejection}$, *scmapcell*, and *scPred* assigned 1.5%, 4.2%, and 10.8% of the cells, respectively, as unlabeled while *SVM* (without rejection) classified 100% of the cells with a median F1-score of 0.98 (Figure 4.1B). This shows an overall better performance for *SVM* and $SVM_{rejection}$, with higher performance and less unlabeled cells.

The CellBench 10X and CEL-Seq2 datasets represent an easy classification task, where the five sorted lung cancer cell lines are quite separable[34]. All classifiers have an almost perfect performance on both CellBench datasets (median F1-score ≈ 1).

For the TM dataset, the top five performing classifiers are $SVM_{rejection}$, *SVM*, *scmapcell*, *Cell-BLAST*, and *scPred* with a median F1-score > 0.96, showing that these classifiers can perform well and scale to large scRNA-seq datasets with a deep level of annotation. Furthermore, *scmapcell* and *scPred* assigned 9.5% and 17.7% of the cells, respectively, as unlabeled, which shows a superior performance for $SVM_{rejection}$ and *SVM*, with a higher median F1-score and 2.9% and 0% unlabeled cells, respectively.

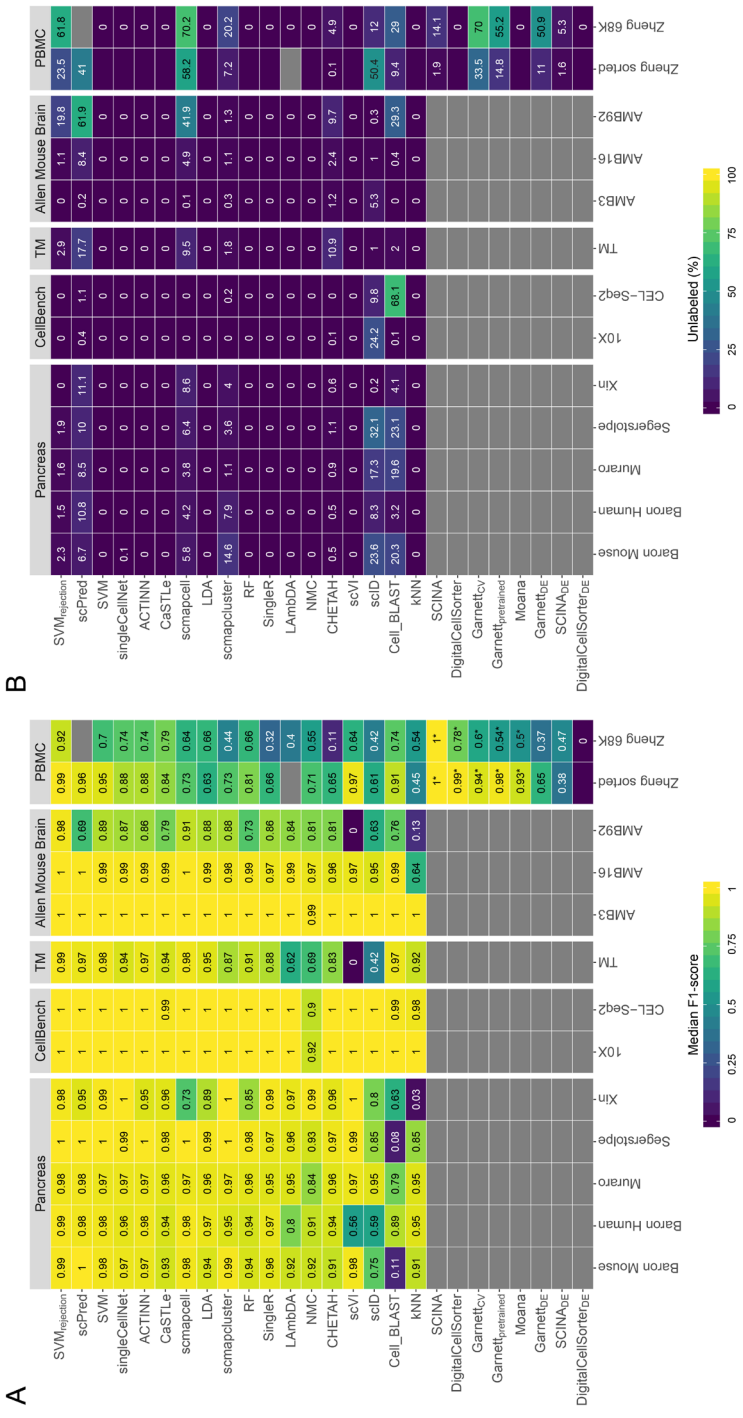


Figure 4.1 Performance comparison of supervised classifiers for cell identification using different scRNA-seq datasets. Heatmap of the **(A)** median F1-scores and **(B)** percentage of unlabeled

cells across all cell populations per classifier (rows) per dataset (columns). Grey boxes indicate that the corresponding method could not be tested on the corresponding dataset. Classifiers are ordered based on the mean of the median F1-scores. Asterix (*) indicates that the prior-knowledge classifiers, *SCINA*, *DigitalCellSorter*, $Garnett_{CV}$, $Garnett_{pretrained}$, and *Moana*, could not be tested on all cell populations of the PBMC datasets. $SCINA_{DE}$, $Garnett_{DE}$, and $DigitalCellSorter_{DE}$ are the versions of *SCINA*, $Garnett_{CV}$, and *DigitalCellSorter* were the marker-genes are defined using differential expression from the training data. Different numbers of marker-genes, 5, 10, 15, and 20, were tested and the best result is shown here. *SCINA, Garnett,* and *DigitalCellSorter* produced the best result for the Zheng sorted dataset using 20, 15 and 5 markers, and for the Zheng 68K dataset using 10, 5 and 5 markers, respectively.

4.2.3 PERFORMANCE EVALUATION ACROSS DIFFERENT ANNOTATION LEVELS

We used the AMB dataset with its three different levels of annotations, to evaluate the classifiers' performance behavior with an increasing number of smaller cell populations within the same dataset. For AMB3, the classification task is relatively easy, differentiating between three major brain cell types (inhibitory neurons, excitatory neurons, and non-neuronal). All classifiers perform almost perfectly with a median F1-score > 0.99 (Figure 4.1A). For AMB16, the classification task becomes slightly more challenging and the performance of some classifiers drops, especially *kNN*. The top five classifiers are $SVM_{rejection}$, *scmapcell*, *scPred*, *SVM*, and *ACTINN*, where $SVM_{rejection}$, *scmapcell*, and *scPred* assigned 1.1%, 4.9%, and 8.4% of the cells as unlabeled, respectively. For the deeply annotated AMB92 dataset, the performance of all classifiers drops further, specially for *kNN* and *scVI*, where the median F1-score is 0.130 and zero, respectively. The top five classifiers are $SVM_{rejection}$, *scmapcell*, *SVM*, *LDA*, and *scmapcluster*, with $SVM_{rejection}$ assigning less cells as unlabeled compared to *scmapcell* (19.8% vs 41.9%), and once more, $SVM_{rejection}$ shows improved performance over *scmapcell* (median F1-score of 0.981 vs 0.906). These results show an overall superior performance for general-purpose classifiers ($SVM_{rejection}$, *SVM*, and *LDA*) compared to other scRNA-seq-specific classifiers across different levels of cell population annotation.

Instead of only looking at the median F1-score, we also evaluated the F1-score per cell population for each classifier (Supplementary Figure 4.1). We confirmed previous conclusions that *kNN* performance drops with deep annotations which include smaller cell populations (Supplementary Figure 4.1B-C), and *scVI* poorly performs on the deeply annotated AMB92 dataset. Additionally, we observed that some cell populations are much harder to classify compared to other populations. For example, most classifiers had a low performance on the *Serpinf1* cells in the AMB16 dataset.

4.2.4 INCORPORATING PRIOR-KNOWLEDGE DOES NOT IMPROVE INTRA-DATASET PERFORMANCE ON PBMCE DATA

For the two PBMC datasets (Zheng 68K and Zheng sorted), the prior-knowledge classifiers *Garnett*, *Moana*, *DigitalCellSorter*, and *SCINA* could be evaluated and benchmarked with the rest of the classifiers. Although the best-performing classifier on Zheng 68K is *SCINA* with a median F1-score of 0.998, this performance is based only on 3, out of 11, cell populations (Monocytes, B cells, and NK cells) for which marker genes are provided. Supplementary Table 4.1 summarizes which PBMC cell populations can be classified by the prior-knowledge methods. Interestingly, none of the prior-knowledge methods showed superior performance compared to other classifiers, despite the advantage these classifiers have over other classifiers given they are tested on fewer cell populations due to the limited availability of marker genes. *Garnett*, *Moana*, and *DigitalCellSorter* could be tested on 7, 7, and 5 cell populations, respectively (Supplementary Table 4.1). Besides *SCINA*, the top classifiers for the Zheng 68K dataset are *CaSTLe*, *ACTINN*, *singleCellNet*, and *SVM*. $SVM_{rejection}$ and *Cell-BLAST* show high performance, at the expense of a high rejection rate of 61.8% and 29%,

respectively (Figure 4.1). Moreover, *scPred* failed when tested on the Zheng 68K dataset. Generally, all classifiers show relatively lower performance on the Zheng 68K dataset compared to other datasets, as the Zheng 68K dataset contains 11 immune cell populations which are harder to differentiate, particularly the T cell compartment (6 out of 11 cell populations). This difficulty of separating these populations was previously noted in the original study[36]. Also, the confusion matrices for *CaSTLe*, *ACTINN*, *singleCellNet*, and *SVM* clearly indicate the high similarity between cell populations, such as (1) monocytes with dendritic cells, (2) the 2 CD8+ T populations, and (3) the 4 CD4+ T populations (Supplementary Figure 4.2).

The classification of the Zheng sorted dataset is relatively easier compared to the Zheng 68K dataset, as almost all classifiers show improved performance (Figure 4.1), with the exception that *LAmbDA* failed while being tested on the Zheng sorted dataset. The prior-knowledge methods show high performance (median F1-score > 0.93), which is still comparable to other classifiers such as $SVM_{rejection}$, *scVI*, *scPred*, and *SVM*. Yet, the supervised classifiers do not require any marker genes, and they can predict more (all) cell populations.

4.2.5 THE PERFORMANCE OF PRIOR-KNOWLEDGE CLASSIFIERS STRONGLY DEPENDS ON THE SELECTED MARKER GENES

Some prior-knowledge classifiers, *SCINA*, *DigitalCellSorter*, and $Garnett_{CV}$, used marker genes to classify the cells. For the PBMC datasets, the number of marker genes per cell population varies across classifiers (2–161 markers) and the marker genes show very little overlap. Only one B cell marker gene, CD79A, is shared by all classifiers while none of the marker genes for the other cell populations is shared by the three classifiers. We analyzed the effect of the number of marker genes, mean expression, dropout rate, and the specificity of each marker gene (beta score, see the "Methods" section) on the performance of the classifier (Supplementary Figure 4.3). The dropout rate and marker specificity (beta-score) are strongly correlated with the median F1-score, highlighting that the performance does not only depend on biological knowledge, but also on technical factors.

The difference between the marker genes used by each method underscores the challenge of marker gene selection, especially for smaller cell populations. Moreover, public databases of cell type markers (e.g., PanglaoDB[39] and CellMarker[40]) often provide different markers for the same population. For example, CellMarker provides 33 marker genes for B cells, while PanglaoDB provides 110 markers, with only 11 marker genes overlap between the two databases.

Given the differences between "expert-defined" markers and the correlation of classification performance and technical dataset-specific features (e.g., dropout rate), we tested if the performance of prior-knowledge methods can be improved by automatically selecting marker genes based on differential expression. Through the cross-validation scheme, we used the training folds to select the marker genes of each cell population based on differential expression (see the "Methods" section) and later used these markers to evaluate the classifiers' performance on the testing fold. We tested this approach on the two PBMC datasets, Zheng sorted and Zheng 68K for different numbers of marker genes (5, 10, 15, and 20 markers). In Figure 4.1, the best result across the number of markers for $SCINA_{DE}$, $Garnett_{DE}$, and $DigitalCellSorter_{DE}$ are shown.

The median F1-score obtained using the differential expression-defined markers is significantly lower compared to the original versions of classifiers using the markers defined by the authors. This lower performance is in part due to the low performance on challenging

populations, such as subpopulations of CD4+ and CD8+ T cell populations (F1-score ≤ 0.68) (Supplementary Figure 4.4). These challenging populations are not identified by the original classifiers since the markers provided by the authors only considered annotations at a higher level (Supplementary Table 4.1). For example, the median F1-score of $SCINA_{DE}$ on Zheng sorted is 0.38, compared to a median F1-score of 1.0 for *SCINA* (using the original markers defined by the authors). However, *SCINA* only considers three cell populations: CD14+ monocytes, CD56+ NK cells, and CD19+ B cells. If we only consider these cell populations for $SCINA_{DE}$, this results in a median F1-score of 0.95.

We observed that the optimal number of marker genes varies per classifier and dataset. For the Zheng sorted dataset, the optimal number of markers is 5, 15, and 20 for $DigitalCellSorter_{DE}$, $Garnett_{DE}$, and $SCINA_{DE}$, respectively, while for Zheng 68K, this is 5, 5, and 10. All together, these results illustrate the dependence of the classification performance on the careful selection of marker genes which is evidently a challenging task.

4.2.6 CLASSIFICATION PERFORMANCE DEPENDS ON DATASET COMPLEXITY

A major aspect affecting the classification performance is the complexity of the dataset at hand. We described the complexity of each dataset in terms of the pairwise similarity between cell populations (see the "Methods" section) and compared the complexity to the performance of the classifiers and the number of cell populations in a dataset (Figure 4.2). When the complexity and/or the number of cell populations of the dataset increases, the performance generally decreases. The performance of all classifiers is relatively low on the Zheng 68K dataset, which can be explained by the high pairwise correlations between the mean expression profiles of each cell population (Supplementary Figure 4.5). These correlations are significantly lower for the TM and AMB92 datasets, justifying the higher performance of the classifiers on these two datasets (Supplementary Figures 4.6–4.7). While both TM and AMB92 have more cell populations (55 and 92, respectively) compared to Zheng 68K (11 populations), these populations are less correlated to one another, making the task easier for all the classifiers.

4.2.7 PERFORMANCE EVALUATION ACROSS DATASETS (INTER-DATASET EVALUATION)

While evaluating the classification performance within a dataset (intra-dataset) is important, the realistic scenario in which a classifier is useful requires cross-dataset (i.e., inter-dataset) classification. We used 22 datasets (Table 4.2) to test the classifiers' ability to predict cell identities in a dataset that was not used for training. First, we tested the classifiers' performance across different sequencing protocols, applied to the same samples within the same lab using the two CellBench datasets. We evaluated the classification performance when training on one protocol and testing on the other. Similar to the intra-dataset evaluation result, all classifiers performed well in this case (Supplementary Figure 4.8).

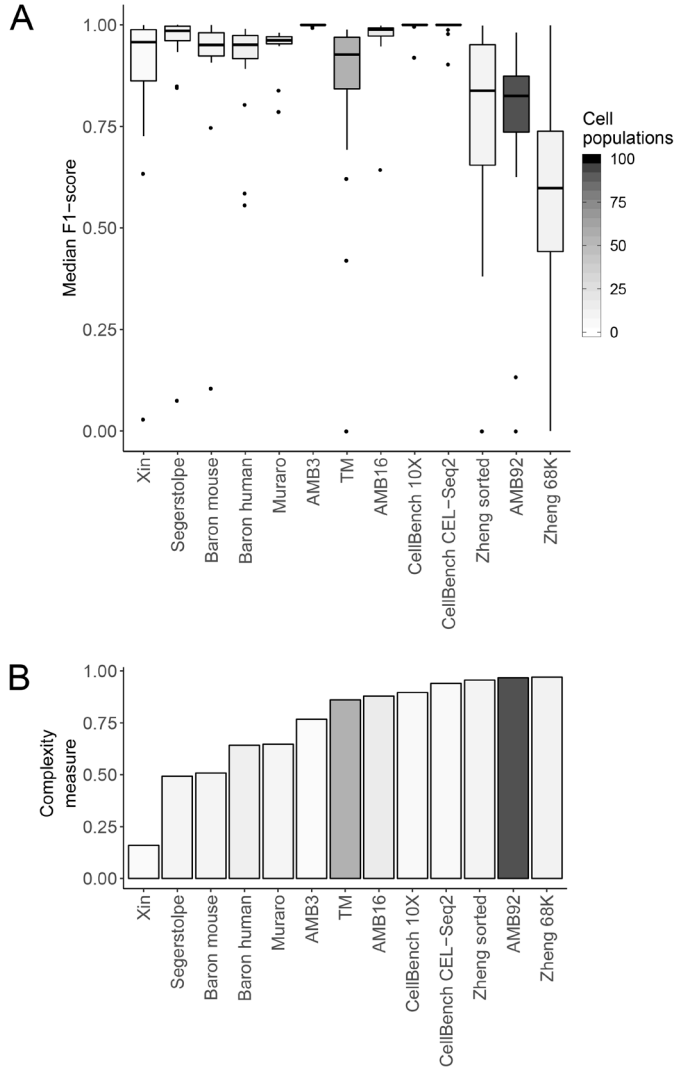


Figure 4.2 Complexity of the datasets compared to the performance of the classifiers. (A) Boxplots of the median F1-scores of all classifiers for each dataset used during the intra-dataset evaluation. **(B)** Barplots describing the complexity of the datasets (see the "Methods" section). Datasets are ordered based on complexity. Box- and barplots are colored according to the number of cell populations in each dataset.

Second, we tested the classification performance on the PbmcBench datasets, which represent a more extensive protocol comparison. PbmcBench consists of 2 samples (pbmc1 and pbmc2), sequenced using 7 different protocols (Table 4.2) with the exception that 10Xv3 was not applied to the pbmc2 sample. We used the pbmc1 datasets to evaluate the classification performance of all pairwise train-test combinations between the 7 protocols (42 experiments, see the "Methods" section). Moreover, we extended the evaluation to include comparisons across different samples for the same protocol, using pbmc1 and pbmc2 (6 experiments, see the "Methods" section). All 48 experiment results are summarized in Figure 4.3. Overall, several classifiers performed well including $SCINA_{DE}$ using 20 marker genes, *singleCellNet*, *scmapcell*, *scID*, and *SVM*, with an average median F1-score > 0.75 across all 48 experiments (Figure 4.3A, Supplementary Figure 4.9A). $SCINA_{DE}$, $Garnett_{DE}$, and $DigitalCellSorter_{DE}$ were tested using 5, 10, 15, and 20 marker genes; Figure 4.3A shows the best result for each classifier, where $SCINA_{DE}$ and $Garnett_{DE}$ performed best using 20 and 5 marker genes, respectively, while $DigitalCellSorter_{DE}$ had a median F1-score of 0 during all experiments using all different numbers of marker genes. $DigitalCellSorter_{DE}$ could only identify B cells in the test sets, usually with an F1-score between 0.8 and 1.0, while the F1-score for all other cell populations was 0.

We also tested the prior-knowledge classifiers on all 13 PbmcBench datasets. The prior-knowledge classifiers showed lower performance compared to other classifiers (average median F1-score < 0.6), with the exception of *SCINA* which was only tested on three cell populations (Figure 4.3B, Supplementary Figure 4.9B). These results are in line with our previous conclusions from the Zheng sorted and Zheng 68K datasets in the intra-dataset evaluation.

Comparing the performance of the classifiers across the different protocols, we observed a higher performance for all classifiers for specific pairs of protocols. For example, all classifiers performed well when trained on 10Xv2 and tested on 10Xv3, and vice versa. On the other hand, other pairs of protocols had a good performance only in one direction, training on Seq-Well produced good predictions on 10Xv3, but not the other way around. Compared to all other protocols, the performance of all classifiers was low when they were either trained or tested on Smart-seq2 data. This can, in part, be due to the fact that Smart-seq2 data does not contain unique molecular identifier (UMI), in contrast to all other protocols.

We also tested the classification performance using the 3 brain datasets, VISp, ALM, and MTG (Table 4.2), which allowed us to compare the performances across species (mouse and human) as well as single-cell RNA-seq (used in VISp and ALM) vs single-nucleus RNA-seq (used in MTG). We tested all possible train-test combinations for both levels of annotation, three major brain cell types (inhibitory neurons, excitatory neurons, and non-neuronal cells), and the deeper annotation level with 34 cell populations (18 experiments, see the "Methods" section). Prediction of the three major cell types was easy, where almost all classifiers showed high performance (Figure 4.4A) with some exceptions. For example, *scPred* failed the classification task completely when testing on the MTG dataset, producing 100% unlabeled cells (Supplementary Figure 4.10A). Predicting the 34 cell populations turned out to be a more challenging task, especially when the MTG human dataset is included either as training or testing data, resulting in significantly lower performance across all classifiers (Figure 4.4B). Across all nine experiments at the deeper annotation, the top-performing classifiers were *SVM*, *ACTINN*, *singleCellNet*, *SingleR*, and *LAmbDA*, with almost 0% unlabeled cells (Supplementary Figure 4.10B).

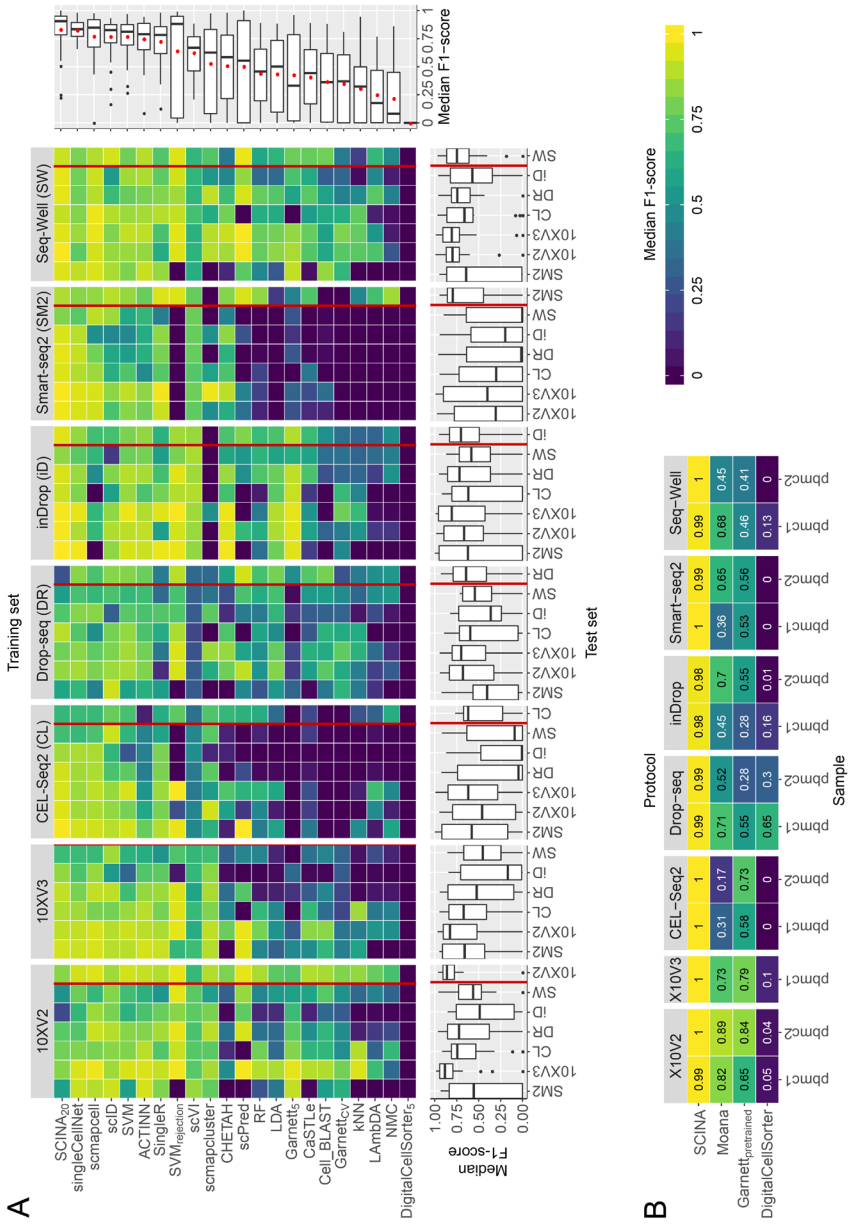


Figure 4.3 Classification performance across the PbmcBench datasets. (A) Heatmap showing the median F1-scores of the supervised classifiers for all train-test pairwise combination across different protocols. The training set is indicated in the grey box on top of the heatmap, the test set is indicated using the column labels below. Results showed to the left of the red line represent the comparison between different protocol using sample pbmc1. Sample pbmc2 was used as test set then. Results showed to the right of the red line represent the comparison between different samples using the same protocol, with pbmc1 used for training and pbmc2 used for testing. Boxplots on the right side of the heatmap summarize the performance of each classifier across all experiments. The mean of the median F1-scores, also used to order the classifiers, is indicated in the boxplots using a red dot. Boxplots underneath the heatmap summarize the performance of the classifiers per experiment. For *SCINA$_{DE}$*, *Garnett$_{DE}$*, and

DigitalCellSorter$_{DE}$ different numbers of marker-genes were tested. Only the best result is shown here. **(B)** Median F1-score of the prior-knowledge classifiers on both samples of the different protocols. The protocol is indicated in the grey box on top of the heatmap, the sample is indicated with the labels below. Classifiers are ordered based on their mean performance across all datasets.

Finally, to evaluate the classification performance across different protocols and different labs, we used the four human pancreatic datasets: Baron Human, Muraro, Segerstople, and Xin (see the "Methods" section, Supplementary Table 4.2). We tested four combinations by training on three datasets and test on one dataset, in which case the classification performance can be affected by batch differences between the datasets. We evaluated the performance of the classifiers when trained using the original data as well as aligned data using the mutual nearest neighbor (MNN) method[41]. Supplementary Figure 4.11 shows UMAPs[42] of the combined dataset before and after alignment, demonstrating better grouping of pancreatic cell types after alignment.

For the original (unaligned) data, the best-performing classifiers across all four experiments are *scVI*, *SVM*, *ACTINN*, *scmapcell*, and *SingleR* (Figure 4.5A, Supplementary Figure 4.12A). For the aligned data, the best-performing classifiers are *kNN*, *SVM*$_{rejection}$, *singleCellNet*, *SVM*, and *NMC* (Figure 4.5B, Supplementary Figure 4.12B). Some classifiers benefit from aligning datasets such as *SVM*$_{rejection}$, *kNN*, *NMC*, and *singleCellNet*, resulting in higher median F1-scores (Figure 4.5). On the other hand, some other classifiers failed the classification task completely, such as *scmapcell* which labels all cells as unlabeled. Some other classifiers failed to run over the aligned datasets, such as *ACTINN*, *scVI*, *Cell-BLAST*, *scID*, *scmapcluster*, and *scPred*. These classifiers work only with positive gene expression data, while the aligned datasets contain positive and negative gene expression values.

4.2.8 REJECTION OPTION EVALUATION

Classifiers developed for scRNA-seq data often incorporate a rejection option to identify cell populations in the test set that were not seen during training. These populations cannot be predicted correctly and therefore should remain unassigned. To test whether the classifiers indeed leave these unseen populations unlabeled, we applied two different experiments using negative controls of different tissues and using unseen populations of the same tissue.

First, the classifiers were trained on a data set from one tissue (e.g., pancreas) and used to predict cell populations of a completely different tissue (e.g., brain)[22]. The methods should thus reject all (100%) of the cells in the test dataset. We carried out four different negative control experiments (see the "Methods" section, Figure 4.6). *scmapcluster* and *scPred* have an almost perfect score for all four combinations, rejecting close 100% of the cells. Other top-performing methods for this task, *SVM*$_{rejection}$ and *scmapcell*, failed when trained on mouse pancreatic data and tested on mouse brain data. All labeled cells of the AMB16 dataset are predicted to be beta cells in this case. The prior-knowledge classifiers, *SCINA*, *Garnett*$_{pretrained}$, and *DigitalCellSorter*, could only be tested on the Baron Human pancreatic dataset. *Garnett*$_{CV}$ could, on top of that, also be trained on the Baron Human dataset and tested on the Zheng 68K dataset. During the training phase, *Garnett*$_{CV}$ tries to find representative cells for the cell populations described in the marker gene file. Being trained on Baron Human using the PBMC marker gene file, it should not be able to find any representatives, and therefore, all cells in the Zheng 68K dataset should be unassigned. Surprisingly, *Garnett*$_{CV}$ still finds representatives for PBMC cells in the pancreatic data, and thus, the cells in the test set are labeled. However, being trained on the PBMC dataset and tested on the pancreatic dataset, it does have a perfect performance.

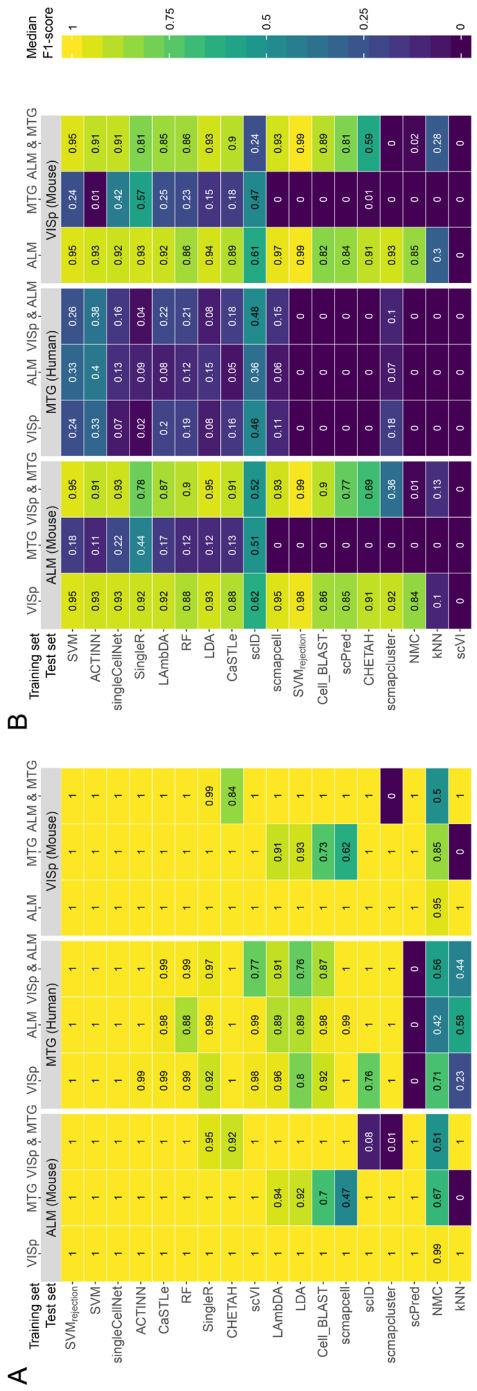


Figure 4.4 Classification performance across brain datasets. Heatmaps show the median F1-scores of the supervised classifiers when tested on **(A)** major lineage annotation with three cell populations, and **(B)** deeper level of annotation with 34 cell populations. The training set(s) are indicated using the column labels on top of the heatmap. The test set is indicated in the grey box. In each heatmap the classifiers are ordered based on their mean performance across all experiments.

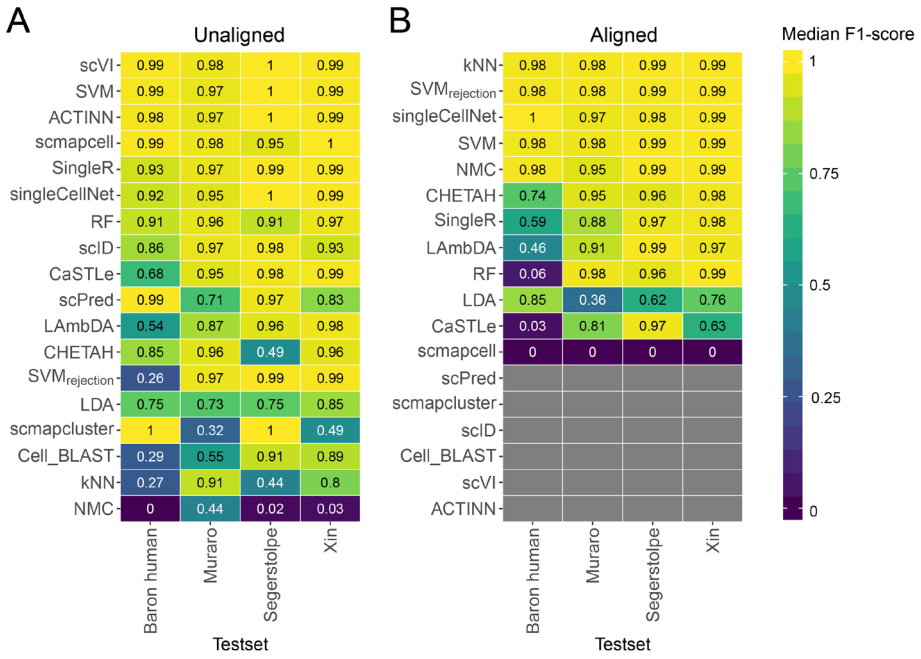


Figure 4.5 Classification performance across pancreatic datasets. Heatmaps showing the median F1-score for each classifier for the **(A)** unaligned and **(B)** aligned datasets. The column labels indicate which of the four datasets was used as a test set, in which case the other three datasets were used as training. Grey boxes indicate that the corresponding method could not be tested on the corresponding dataset. In each heatmap, the classifiers are ordered based on their mean performance across all experiments.

To test the rejection option in a more realistic and challenging scenario, we trained the classifiers on some cell populations from one dataset and used the held out cell populations in the test set (see the "Methods" section). Since the cell populations in the test set were not seen during training, they should remain unlabeled. Here, the difficulty of the task was gradually increased (Supplementary Table 4.3). First, all the T cells were removed from the training set. Next, only the CD4+ T cells were removed. Finally, only CD4+/CD45RO+ memory T cells, a subpopulation of the CD4+ T cells, were removed. The top-performing methods for this task are *scmapcell*, *scPred*, *scID*, $SVM_{rejection}$, and *SCINA* (Figure 4.6B). We expected that rejecting T cells would be a relatively easy task as they are quite distinct from all other cell populations in the dataset. It should thus be comparable to the negative control experiment. Rejecting CD4+/CD45RO+ memory T cells, on the other hand, would be more difficult as they could easily be confused with all other subpopulations of CD4+ T cells. Surprisingly, almost all classifiers, except for *scID* and *scmapcluster*, show the opposite.

To better understand this unexpected performance, we analyzed the labels assigned by $SVM_{rejection}$. In the first task (T cells removed from the training set), $SVM_{rejection}$ labels almost all T cells as B cells. This can be explained by the fact that $SVM_{rejection}$, and most classifiers for that matter, relies on the classification posterior probabilities to assign labels but ignores the actual similarity between each cell and the assigned population. In task 2 (CD4+ T cells were removed), there were two subpopulations of CD8+ T cells in the training set. In that case, two cell populations are equally similar to the cells in the test set, resulting in low posterior probabilities for both classes and thus the cells in the test set remain

unlabeled. If one of these CD8+ T cell populations was removed from the training set, only 10.53% instead of 75.57% of the CD4+ T cells were assigned as unlabeled by $SVM_{rejection}$. All together, our results indicate that despite the importance of incorporating a rejection option in cell identity classifiers, the implementation of this rejection option remains challenging.

4.2.9 PERFORMANCE SENSITIVITY TO THE INPUT FEATURES

During the intra-datasets cross-validation experiment described earlier, we used all features (genes) as input to the classifiers. However, some classifiers suffer from overtraining when too many features are used. Therefore, we tested the effect of feature selection on the performance of the classifiers. While different strategies for feature selection in scRNA-seq classification experiments exist, selecting genes with a higher number of dropouts compared to the expected number of dropouts has been shown to outperform other methods[22,43]. We selected subsets of features from the TM dataset using the dropout method. In the experiments, we used the top 100, 200, 500, 1000, 2000, 5000, and 19,791 (all) genes. Some classifiers include a built-in feature selection method which is used by default. To ensure that all methods use the same set of features, the built-in feature selection was turned off during these experiments.

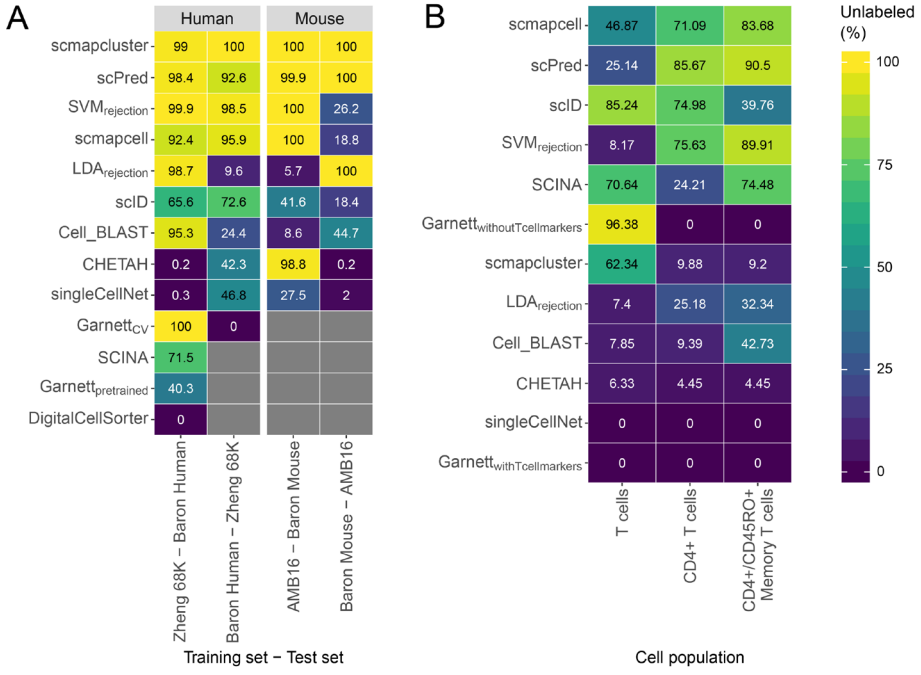


Figure 4.6 Performance of the classifiers during the rejection experiments. (A) Percentage of unlabeled cells during the negative control experiment for all the classifiers with a rejection option. The prior-knowledge classifiers could not be tested on all datasets, this is indicated with a grey box. The species of the dataset is indicated in the grey box on top. Column labels indicate which datasets are used for training and testing respectively. **(B)** Percentage of unlabeled cells for all classifiers with a rejection option when a cell population was removed from the training set. Column labels indicate which cell population was removed. This cell population was used as a test set. In both **(A)** and **(B)** the classifiers are sorted based on their mean performance across all experiments.

Some methods are clearly overtrained when the number of features increases (Figure 4.7A). For example, *scmapcell* shows the highest median F1-score when using less features, and the performance drops when the number of features increases. On the other hand, the performance of other classifiers, such as *SVM*, keeps improving when the number of features increases. These results indicate that the optimal number of features is different for each classifier.

Looking at the median F1-score, there are several methods with a high maximal performance. *Cell-BLAST*, *ACTINN*, *scmapcell*, *scPred*, $SVM_{rejection}$, and *SVM* all have a median F1-score higher than 0.97 for one or more of the feature sets. Some of these well-performing methods, however, leave many cells unlabeled. *Scmapcell* and *scPred*, for instance, yield a maximum median F1-score of 0.976 and 0.982, respectively, but 10.7% and 15.1% of the cells are assigned as unlabeled (Figure 4.7B). On the other hand, $SVM_{rejection}$ has the highest median F1-score (0.991) overall with only 2.9% unlabeled. Of the top-performing classifiers, only *ACTINN* and *SVM* label all the cells. Overall *SVM* shows the third highest performance with a score of 0.979.

4.2.10 SCALABILITY: PERFORMANCE SENSITIVITY TO THE NUMBER OF CELLS

scRNA-seq datasets vary significantly across studies in terms of the number of cells analyzed. To test the influence of the size of the dataset on the performance of the classifier, we downsampled the TM dataset in a stratified way (i.e., preserving population frequencies) to 1, 5, 10, 20, 50, and 100% of the original number of 45,469 cells (see the "Methods" section) and compared the performance of the classifiers (Figure 4.7C-D). Using less than 500 cells in the dataset, most classifiers have a relatively high performance. Only *scID*, *LAmbDA*, *CaSTLe*, and *Cell-BLAST* have a median F1-score below 0.85. Surprisingly, $SVM_{rejection}$ has almost the same median F1-score when using 1% of the data as when using all data (0.993 and 0.994). It must be noted here, however, that the percentage of unlabeled cells decreases significantly (from 28.9% to 1.3%). Overall, the performance of all classifiers stabilized when tested on ≥ 20% (9099 cells) of the original data.

4.2.11 RUNNING TIME EVALUATION

To compare the runtimes of the classification methods and see how they scale when the number of cells increases, we compared the number of cells in each dataset with the computation time of the classifiers (Supplementary Figure 4.13). Overall, big differences in the computation time can be observed when comparing the different methods. *SingleR* showed the highest computation time overall. Running *SingleR* on the Zheng 68K dataset took more than 39 h, while *scmapcluster* was finished within 10 s on this dataset. Some of the methods have a high runtime for the small datasets. On the smallest dataset, Xin, all classifiers have a computation time < 5 min, with most classifiers finishing within 60 s. *Cell-BLAST*, however, takes more than 75 min. In general, all methods show an increase in computation time when the number of cells increases. However, when comparing the second largest (TM) and the largest (Zheng 68K) datasets, not all methods show an increase in computation time. Despite the increase in the number of cells between the two datasets, *CaSTLe*, *CHETAH*, and *SingleR* have a decreasing computation time. A possible explanation could be that the runtime of these methods also depends on the number of genes or the number of cell populations in the dataset. To evaluate the run time of the methods properly, we therefore investigated the effect of the number of cells, features, and cell populations separately (Figure 4.7E-G).

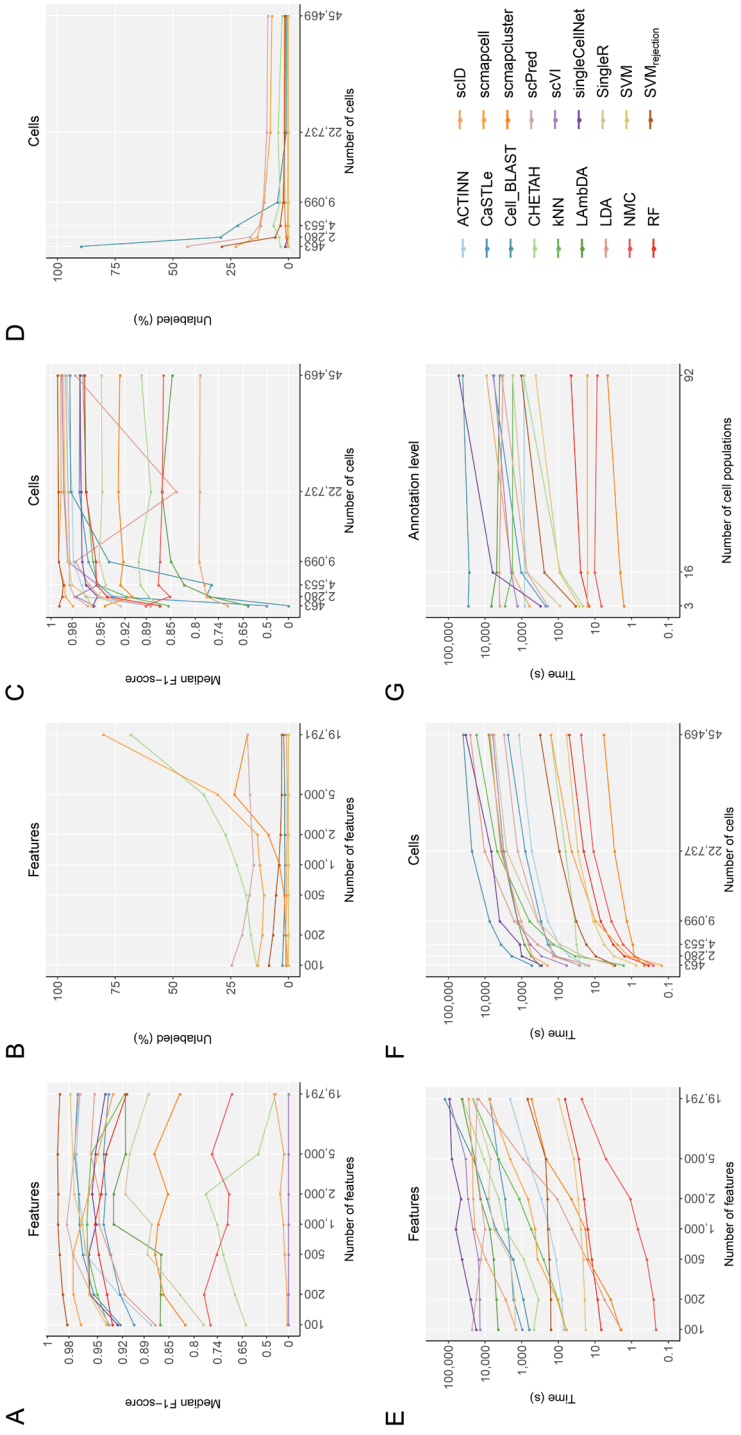


Figure 4.7 Classification performance and computation time evaluation across different numbers of features, cells, and annotation levels. Line plots show **(A)** the median F1-score, **(B)** percentage of unlabeled cells, and **(E)** computation time of each classifier applied to the TM dataset with

the top 100, 200, 500, 1000, 2000, 5000, and 19791 (all) genes as input feature sets. Genes were ranked based on dropout-based feature selection. **(C)** The median F1-score, **(D)** percentage of unlabeled cells, and **(F)** computation time of each classifier applied to the downsampled TM datasets containing 463, 2,280, 4,553, 9,099, 22,737, and 45,469 (all) cells. **(G)** The computation time of each classifier is plotted against the number of cell populations. Note that the y-axis is 100^x scaled in **(A,C)** and log-scaled in **(E-G).** The x-axis is log-scaled in **(A-F)**.

To assess the effect of the number of genes on the computation time, we compared the computation time of the methods during the feature selection experiment (Figure 4.7E). Most methods scale linearly with the number of genes. However, *LDA* does not scale very well when the number of genes increases. If the number of features is higher than the number of cells, the complexity of *LDA* is O(*g*^3), where *g* is the number of genes[44].

The effect of the number of cells on the timing showed that all methods increase in computation time when the number of cells increases (Figure 4.7F). The differences in runtime on the largest dataset are larger. *scmapcluster*, for instance, takes 5 s to finish, while *Cell-BLAST* takes more than 11 h.

Finally, to evaluate the effect of the number of cell populations, the runtime of the methods on the AMB3, AMB16, and AMB92 datasets was compared (Figure 4.7G). For most methods, this shows an increase in runtime when the number of cell populations increases, specially *singleCellNet*. For other methods, such as *ACTINN* and *scmapcell*, the runtime remains constant. Five classifiers, *scmapcell*, *scmapcluster*, *SVM*, *RF*, and *NMC*, have a computation time below 6 min on all the datasets.

4.3 DISCUSSION

In this study, we evaluated the performance of 22 different methods for automatic cell identification using 27 scRNA-seq datasets. We performed several experiments to cover different levels of challenges in the classification task and to test specific aspects of the classifiers such as the feature selection, scalability, and rejection experiments. We summarize our findings across the different experiments (Figure 4.8) and provide a detailed summary of which dataset was used for each experiment (Supplementary Table 4.4). This overview can be used as a user guide to choose the most appropriate classifier depending on the experimental setup at hand. Overall, several classifiers performed accurately across different datasets and experiments, particularly $SVM_{rejection}$, *SVM*, *singleCellNet*, *scmapcell*, *scPred*, *ACTINN*, and *scVI*. We observed relatively lower performance for the inter-dataset setup, likely due to the technical and biological differences between the datasets, compared to the intra-dataset setup. $SVM_{rejection}$, *SVM*, and *singleCellNet* performed well for both setups, while *scPred* and *scmapcell* performed better in the intra-dataset setup, and *scVI* and *ACTINN* had a better performance in the inter-dataset setup (Figure 4.8). Of note, we evaluated all classifiers using the default settings. While adjusting these settings for a specific dataset might improve the performances, it increases the risk of overtraining.

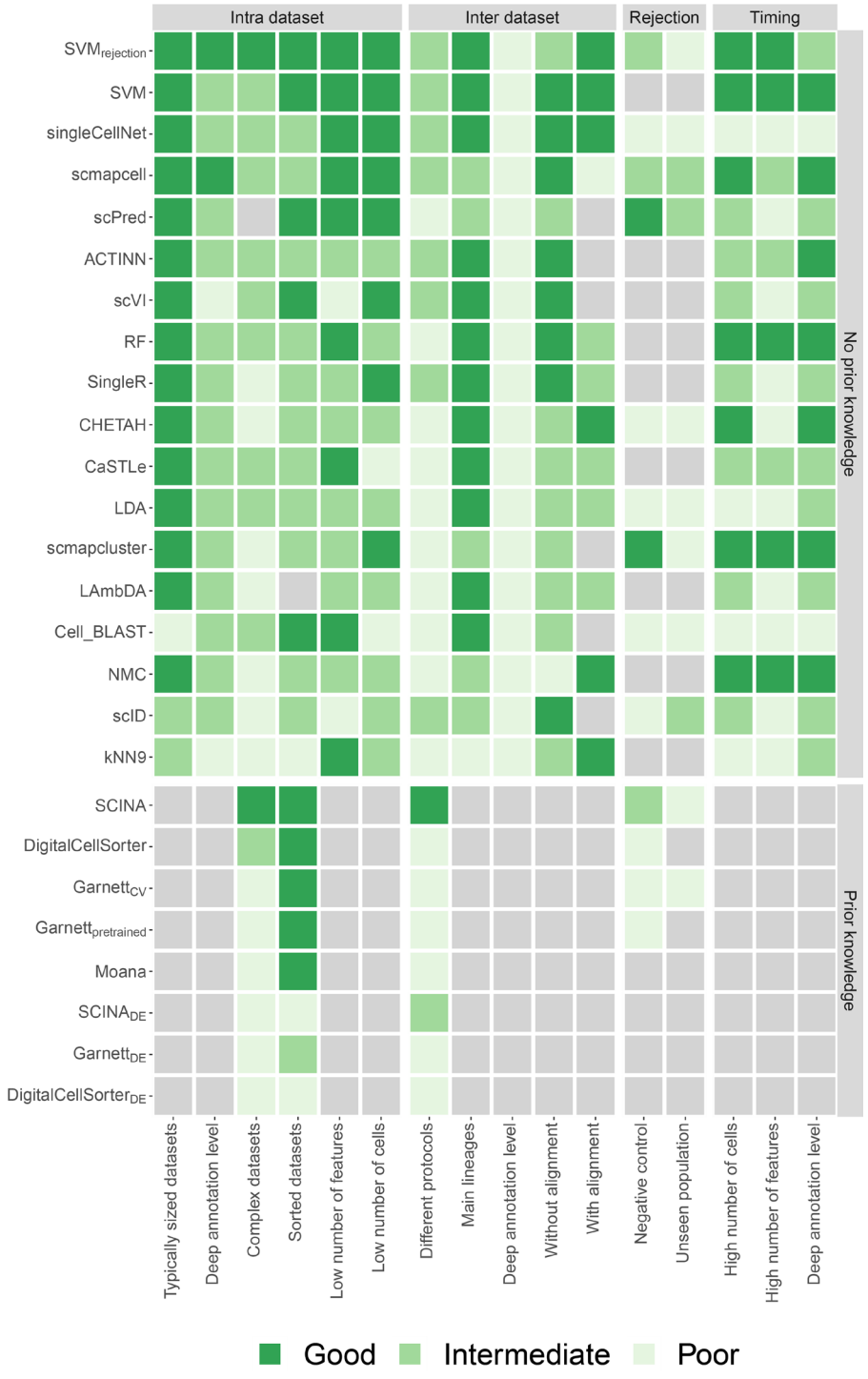


Figure 4.8 Summary of the performance of all classifiers during different experiments. For each experiment, the heatmap shows whether a classifiers performs good, intermediate, or poor. Light-grey

indicates that a classifier could not be tested during an experiment. The grey boxes to the right of the heatmap indicate the four different categories of experiments: intra dataset, inter dataset, rejection and timing. Experiments itself are indicated using the row labels. Supplementary Table 4.4 shows which datasets were used to score the classifiers exactly for each experiment. Grey boxes above the heatmap indicate the two classifiers categories. Within these two categories, the classifiers are sorted based on their mean performance on the intra and inter dataset experiments.

Considering all three evaluation metrics (median F1-score, percentage of unlabeled cells, and computation time), $SVM_{rejection}$ and *SVM* are overall the best-performing classifiers for the scRNA-seq datasets used. Although *SVM* has a shorter computation time, the high accuracy of the rejection option of $SVM_{rejection}$, which allows flagging new cells and assigning them as unlabeled, results in an improved performance compared to *SVM*. Our results show that $SVM_{rejection}$ and *SVM* scale well to large datasets as well as deep annotation levels. In addition, they did not suffer from the large number of features (genes) present in the data, producing the highest performance on the TM dataset using all genes, due to the incorporated L2 regularization. The comparable or higher overall performance of a general-purpose classier such as *SVM* warrants caution when designing scRNA-seq-specific classifiers that they do not introduce unnecessary complexity. For example, deep learning methods, such as *ACTINN* and *scVI*, showed overall lower performance compared to *SVM*, supporting recent observations by Köhler et al.[45].

scPred (which is based on an SVM with a radial kernel), *LDA*, *ACTINN*, and *singleCellNet* performed well on most datasets, yet the computation time is long for large datasets. *singleCellNet* also becomes slower with a large number of cell populations. Additionally, in some cases, *scPred* and *scmapcell/cluster* reject higher proportions of cells as unlabeled compared to $SVM_{rejection}$, without a substantial improvement in the accuracy. In general, incorporating a rejection option with classification is a good practice to allow the detection of potentially novel cell populations (not present in the training data) and improve the performance for the classified cells with high confidence. However, for the datasets used in this study, the performance of classifiers with a rejection option, except for $SVM_{rejection}$, did not show substantial improvement compared to other classifiers. Furthermore, our results indicate that designing a proper rejection option can be challenging for complex datasets (e.g., PBMC) and that relying on the posterior probabilities alone might not yield optimal results.

For datasets with deep levels of annotation (i.e., large number) of cell populations, the classification performance of all classifiers is relatively low, since the classification task is more challenging. *scVI*, in particular, failed to scale with deeply annotated datasets, although it works well for datasets with a relatively small number of cell populations. Further, applying the prior-knowledge classifiers becomes infeasible for deeply annotated datasets, as the task of defining the marker genes becomes even more challenging.

We evaluated the performance of the prior-knowledge methods (marker-based and pretrained) on PBMC datasets only, due to the limited availability of author-provided marker genes. For all PBMC datasets, the prior-knowledge methods did not improve the classification performance over supervised methods, which do not incorporate such prior knowledge. We extended some prior-knowledge methods such that the marker genes were defined in a data-driven manner using differential expression which did not improve the performance of these classifiers, except for $SCINA_{DE}$ (with 20 marker genes) for the PbmcBench datasets. The data-driven selection of markers allows the prediction of more cell populations compared to the number of populations for which marker genes were originally provided. However, this data-driven selection violates the fundamental assumption in prior-knowledge methods that incorporating expert-defined markers improves classification performance. Further, several

supervised classifiers which do not require markers to be defined a priori (e.g., *scPred* and *scID*) already apply a differential expression test to find the best set of genes to use while training the model. The fact that prior-knowledge methods do not outperform other supervised methods and given the challenges associated with explicit marker definition indicate that incorporating prior knowledge in the form of marker genes is not beneficial, at least for PBMC data.

In the inter-dataset experiments, we tested the ability of the classifiers to identify populations across different scRNA-seq protocols. Our results show that some protocols are more compatible with one another (e.g., 10Xv2 and 10Xv3), Smart-Seq2 is distinct from the other UMI-based methods, and CEL-Seq2 suffers from low replicability of cell populations across samples. These results can serve as a guide in order to choose the best set of protocols that can be used in studies where more than one protocol is used.

The intra-dataset evaluation included the Zheng sorted dataset, which consists of 10 FACS-sorted cell populations based on the expression of surface protein markers. Our results show relatively lower classification performance compared to other datasets, except the Zheng 68K dataset. The poor correlation between the expression levels of these protein markers and their coding genes mRNA levels[46] might explain this low performance.

Overall, we observed that the performance of almost all methods was relatively high on various datasets, while some datasets with overlapping populations (e.g., Zheng 68K dataset) remain challenging. The inter-dataset comparison requires extensive development in order to deal with technical differences between protocols, batches, and labs, as well as proper matching between different cell population annotations. Further, the pancreatic datasets are known to project very well across studies, and hence, using them to evaluate inter-dataset performance can be misleading. We recommend considering other challenging tissues and cell populations.

4.4 CONCLUSIONS

We present a comprehensive evaluation of automatic cell identification methods for single-cell RNA sequencing data. Generally, all classifiers perform well across all datasets, including the general-purpose classifiers. In our experiments, incorporating prior knowledge in the form of marker genes does not improve the performance (on PBMC data). We observed large differences in the performance between methods in response to changing the input features. Furthermore, the tested methods vary considerably in their computation time which also varies differently across methods based on the number of cells and features.

Taken together, we recommend the use of the general-purpose $SVM_{rejection}$ classifier (with a linear kernel) since it has a better performance compared to the other classifiers tested across all datasets. Other high-performing classifiers include *SVM* with a remarkably fast computation time at the expense of losing the rejection option, *singleCellNet*, *scmapcell*, and *scPred*. To support the future extension of this benchmarking work with new classifiers and datasets, we provide a Snakemake workflow to automate the performed benchmarking analyses (https://github.com/tabdelaal/scRNAseq_Benchmark/).

4.5 METHODS

4.5.1 CLASSIFICATION METHODS

We evaluated 22 scRNA-seq classifiers, publicly available as R or Python packages or scripts (Table 4.1). This set includes 16 methods developed specifically for scRNA-seq data as well as 6 general-purpose classifiers from the scikit-learn library in Python[29]: linear discriminant analysis (*LDA*), nearest mean classifier (*NMC*), *k*-nearest neighbor (*kNN*), support vector machine (*SVM*) with linear kernel, SVM with rejection option ($SVM_{rejection}$), and random forest (*RF*). The following functions from the scikit-learn library were used respectively: LinearDiscriminantAnalysis(), NearestCentroid(), KNeighborsClassifier(n_neighbors=9), LinearSVC(), LinearSVC() with CalibratedClassifierCV() wrapper, and RandomForestClassifier(n_estimators=50). For *kNN*, 9 neighbors were chosen. After filtering the datasets, only cell populations consisting of 10 cells or more remained. Using 9 neighbors would thus ensure that this classifier could also predict very small populations. For $SVM_{rejection}$, a threshold of 0.7 was used on the posterior probabilities to assign cells as "unlabeled." During the rejection experiments, also an LDA with rejection was implemented. In contrast to the LinearSVC(), the LinearDiscriminantAnalysis() function can output the posterior probabilities, which was also thresholded at 0.7.

scRNA-seq-specific methods were excluded from the evaluation if they did not return the predicted labels for each cell. For example, we excluded *MetaNeighbor*[47] because the tool only returns the area under the receiver operator characteristic curve (AUROC). For all methods, the latest (May 2019) package was installed or scripts were downloaded from their GitHub. For *scPred*, it should be noted that it is only compatible with an older version of Seurat (v2.0). For *CHETAH*, it is important that the R version 3.6 or newer is installed. For *LAmbDA*, instead of the predicted label, the posterior probabilities were returned for each cell population. Here, we assigned the cells to the cell population with the highest posterior probability.

During the benchmark, all methods were run using their default settings, and if not available, we used the settings provided in the accompanying examples or vignettes. As input, we provided each method with the raw count data (after cell and gene filtering as described in the "Data preprocessing" section) according to the method documentation. The majority of the methods have a built-in normalization step. For the general-purpose classifiers, we provided log-transformed counts, $log_2(count + 1)$.

Some methods required a marker gene file or pretrained classifier as an input (e.g., *Garnett*, *Moana*, *SCINA*, *DigitalCellSorter*). In this case, we use the marker gene files or pretrained classifiers provided by the authors. We did not attempt to include additional marker gene files for all datasets, and hence, the evaluation of those methods is restricted to datasets where a marker gene file for cell populations is available.

4.5.2 DATASETS

A total of 27 scRNA-seq datasets were used to evaluate and benchmark all classification methods, from which 11 datasets were used for intra-dataset evaluation using a cross-validation scheme, and 22 datasets were used for inter-dataset evaluation, with 6 datasets overlapping for both tasks as described in Table 4.2. Datasets vary across species (human and mouse), tissue (brain, pancreas, PBMC, and whole mouse), and the sequencing protocol used. The brain datasets, including Allen Mouse Brain (AMB), VISp, ALM (GSE115746), and MTG (phs001790), were downloaded from the Allen Institute Brain Atlas

http://celltypes.brain-map.org/rnaseq. All 5 pancreatic datasets were obtained from https://hemberg-lab.github.io/scRNA.seq.datasets/ (Baron Mouse: GSE84133, Baron Human: GSE84133, Muraro: GSE85241, Segerstolpe: E-MTAB-5061, Xin: GSE81608). The CellBench 10X dataset was obtained from (GSM3618014), and the CellBench CEL-Seq2 dataset was obtained from 3 datasets (GSM3618022, GSM3618023, GSM3618024) and concatenated into 1 dataset. The Tabula Muris (TM) dataset was downloaded from https://tabula-muris.ds.czbiohub.org/ (GSE109774). For the Zheng sorted datasets, we downloaded the 10 PBMC-sorted populations (CD14+ monocytes, CD19+ B cells, CD34+ cells, CD4+ helper T cells, CD4+/CD25+ regulatory T cells, CD4+/CD45RA+/CD25− naive T cells, CD4+/CD45RO+ memory T cells, CD56+ natural killer cells, CD8+ cytotoxic T cells, CD8+/CD45RA+ naive cytotoxic T cells) from https://support.10xgenomics.com/single-cell-gene-expression/datasets; next, we downsampled each population to 2000 cells obtaining a dataset of 20,000 cells in total. For the Zheng 68K dataset, we downloaded the gene-cell count matrix for the "Fresh 68K PBMCs"[36] from https://support.10xgenomics.com/single-cell-gene-expression/datasets (SRP073767). All 13 PbmcBench datasets, 7 different sequencing protocols applied on 2 PBMC samples, were downloaded from the Broad Institute Single Cell portal https://portals.broadinstitute.org/single_cell/study/SCP424/single-cell-comparison-pbmc-data. The cell population annotation for all datasets was provided with the data, except the Zheng 68K dataset, for which we obtained the cell population annotation from https://github.com/10XGenomics/single-cell-3prime-paper/tree/master/pbmc68k_analysis. These annotations were used as a "ground truth" during the evaluation of the cell population predictions obtained from the classification methods.

4.5.3 DATA PREPROCESSING

Based on the manual annotation provided in the datasets, we started by filtering out cells that were labeled as doublets, debris, or unlabeled cells. Next, we filtered genes with zero counts across all cells. For cells, we calculated the median number of detected genes per cell, and from that, we obtained the median absolute deviation (MAD) across all cells in the log scale. We filtered out cells when the total number of detected genes was below three MAD from the median number of detected genes per cell. The number of cells and genes in Table 4.2 represent the size of each dataset after this stage of preprocessing.

Moreover, before applying cross-validation to evaluate each classifier, we excluded cell populations with less than 10 cells across the entire dataset; Table 4.2 summarizes the number of cell populations before and after this filtration step for each dataset.

4.5.4 INTRA-DATASET CLASSIFICATION

For the supervised classifiers, we evaluated the performance by applying a 5-fold cross-validation across each dataset after filtering genes, cells, and small cell populations. The folds were divided in a stratified manner in order to keep equal proportions of each cell population in each fold. The training and testing folds were exactly the same for all classifiers.

The prior-knowledge classifiers, *Garnett*, *Moana*, *DigitalCellSorter*, and *SCINA*, were only evaluated on the Zheng 68K and Zheng sorted datasets, for which the marker gene files or the pretrained classifiers were available, after filtering genes and cells. Each classifier uses the dataset and the marker gene file as inputs and outputs the cell population label corresponding to each cell. No cross-validation is applied in this case, except for *Garnett* where we could either use the pretrained version ($Garnett_{pretrained}$) provided from the original study, or train our own classifier using the marker gene file along with the training data ($Garnett_{CV}$). In this case, we applied 5-fold cross-validation using the same train and test sets described earlier. Supplementary Table 4.1 shows the mapping of cell

populations between the Zheng datasets and each of the prior-knowledge classifiers. For *Moana*, a pretrained classifier was used, this classifier also predicted cells to be memory CD8+ T cells and CD16+ monocytes, while these cell populations were not in the Zheng datasets.

4.5.5 EVALUATION OF MARKER GENES

The performance and choice of the marker genes per cell population per classifier were evaluated by comparing the F1-score of each cell population with four different characteristics of the marker genes across the cells for that particular cell population: (1) the number of marker genes, (2) the mean expression, (3) the average dropout rate, and (4) the average beta of the marker genes[37]. Beta is a score developed to measure how specific a marker gene for a certain cell population is based on binary expression.

4.5.6 SELECTING MARKER GENES USING DIFFERENTIAL EXPRESSION

Using the cross-validation scheme, training data of each fold was used to select sets of 5, 10, 15, and 20 differentially expressed (DE) marker genes. First, if the data was not already normalized, a CPM read count normalization was applied to the data. Next, the data was log-transformed using $log_2(count + 1)$, and afterwards, the DE test could be applied. As recommended in[48], MAST was used to find the DE genes[49]. The implementation of MAST in the FindAllMarkers() function of Seurat v2.3.0 was used to do a one-vs-all differential expression analysis[50]. Genes returned by Seurat were sorted, and the top 5, 10, 15, or 20 significant genes with a positive fold change were selected as marker genes. These marker genes were then used for population prediction of the test data of the corresponding fold. These marker gene lists can be used by prior-knowledge classifiers such as *SCINA*, $Garnett_{CV}$, and *DigitalCellSorter*, by modifying the cell type marker gene file required as an input to these classifiers. Such modification cannot be applied to the pretrained classifiers of $Garnett_{pretrained}$ and *Moana*.

4.5.7 DATASET COMPLEXITY

To describe the complexity of a dataset, the average expression of all genes for each cell population (avg_{C_i}) in the dataset was calculated, representing the prototype of each cell population in the full genes space. Next, the pairwise Pearson correlation between these centroids was calculated $\underset{\forall\, i,j}{\text{corr}}(avg_{C_i}, avg_{C_j})$. For each cell population, the highest correlation to another cell population was recorded. Finally, the mean of these per cell population maximum correlations was taken to describe the complexity of a dataset.

$$Complexity = mean(\max_{\forall\, i,l \neq j} \underset{\forall\, i,j}{\text{corr}}(avg_{C_i}, avg_{C_j}))$$

4.5.8 INTER-DATASET CLASSIFICATION

4.5.8.1 CellBench

Both CellBench datasets, 10X and CEL-Seq2, were used once as training data and once as test data, to obtain predictions for the five lung cancer cell lines. The common set of detected genes by both datasets was used as features in this experiment.

4.5.8.2 PbmcBench

Using pbmc1 sample only, we tested all train-test pairwise combinations between all 7 protocols, resulting in 42 experiments. Using both pbmc1 and pbmc2 samples, for the same

protocol, we used pbmc1 as training data and pbmc2 as test data, resulting in 6 additional experiments (10Xv3 was not applied for pbmc2). As we are now dealing with PBMC data, we evaluated all classifiers, including the prior-knowledge classifiers, as well as the modified versions of *SCINA*, *Garnett*$_{CV}$, and *DigitalCellSorter*, in which the marker genes are obtained through differential expression from the training data as previously described. Through all these 48 experiments, genes that are not expressed in the training data were excluded from the feature space. Also, as these PbmcBench datasets differ in the number of cell populations (Table 4.2), only the cell populations provided by the training data were used for the test data prediction evaluation.

4.5.8.3 Brain

We used the three brain datasets, VISp, ALM, and MTG with two levels of annotations, 3 and 34 cell populations. We tested all possible train-test combinations, by either using one dataset to train and test on another (6 experiments) or using two concatenated datasets to train and test on the third (3 experiments). A total of 9 experiments were applied for each annotation level. We used the common set of detected genes between the datasets involved in each experiment as features.

4.5.8.4 Pancreas

We selected the four major endocrine pancreatic cell types (alpha, beta, delta, and gamma) across all four human pancreatic datasets: Baron Human, Muraro, Segerstolpe, and Xin. Supplementary Table 4.2 summarizes the number of cells in each cell type across all datasets. To account for batch effects and technical variations between different protocols, datasets were aligned using MNN[41] from the scran R package (version 1.1.2.0). Using both the raw data (unaligned) and the aligned data, we applied leave-one-dataset-out cross-validation where we train on three datasets and test on the left out dataset.

4.5.9 PERFORMANCE EVALUATION METRICS

The performance of the methods on the datasets is evaluated using three different metrics: (1) For each cell population in the dataset, the F1-score is reported. The median of these F1-scores is used as a measure for the performance on the dataset. (2) Some of the methods do not label all the cells. These unassigned cells are not considered in the F1-score calculation. The percentage of unlabeled cells is also used to evaluate the performance. (3) The computation time of the methods is also measured.

4.5.10 FEATURE SELECTION

Genes are selected as features based on their dropout rate. The method used here is based on the method described in[22]. During feature selection, a sorted list of the genes is made. Based on this list, the top n number of genes can be easily selected during the experiments. First, the data is normalized using $log_2(count + 1)$. Next, for each gene, the percentage of dropouts, d, and the mean, m, of the normalized data are calculated. Genes that have a mean or dropout rate of 0 are not considered during the next steps. These genes will be at the bottom of the sorted list. For all other genes, a linear model is fitted to the mean and log2(d). Based on their residuals, the genes are sorted in descending order and added to the top of the list.

4.5.11 SCALABILITY

For the scalability experiment, we used the TM dataset. To ensure that the dataset could be downsampled without losing cell populations, only the 16 most abundant cell populations

were considered during this experiment. We downsampled these cell populations in a stratified way to 1, 5, 10, 20, 50, and 100% of its original size (45,469 cells).

4.5.12 REJECTION

4.5.12.1 Negative Control

Two human datasets, Zheng 68K and Baron Human, and two mouse datasets, AMB16 and Baron Mouse, were used. The Zheng 68K dataset was first stratified downsampled to 11% of its original size to reduce computation time. For each species, two different experiments were applied by using one dataset as a training set and the other as a test set and vice versa.

4.5.12.2 Unseen Cell Populations

Zheng 68K dataset was stratified downsampled to 11% of its original size to reduce computation time. Three different experiments were conducted. First, all cell populations that are a subpopulation of T cells were considered the test set. Next, the test set consisted of all subpopulations of CD4+ T cells. Last, only the CD4+/CD45RO+ memory T cells were in the test set. Each time, all cell populations that were not in the test set were part of the training set. Supplementary Table 4.3 gives an exact overview of the populations per training and test set.

4.5.13 BENCHMARKING PIPELINE

In order to ensure reproducibility and support the future extension of this benchmarking work with new classification methods and benchmarking datasets, a Snakemake[51] workflow for automating the performed benchmarking analyses was developed with an MIT license (https://github.com/tabdelaal/scRNAseq_Benchmark/). Each tool (license permitting) is packaged in a Docker container (https://hub.docker.com/u/scrnaseqbenchmark) alongside the wrapper scripts and their dependencies. These images will be used through Snakemake's singularity integration to allow the workflow to be run without the requirement to install specific methods and to ensure reproducibility. Documentation is also provided to execute and extend this benchmarking workflow to help researchers to further evaluate interested methods.

4.6 AVAILABILITY OF DATA AND MATERIALS

The filtered datasets analyzed during the current study can be downloaded from Zenodo (https://doi.org/10.5281/zenodo.3357167). The source code is available in the GitHub repository, at https://github.com/tabdelaal/scRNAseq_Benchmark, and in the Zenodo repository, at https://doi.org/10.5281/zenodo.3369158. The source code is released under MIT license. Datasets accession numbers: AMB, VISp, and ALM[35] (GSE115746), MTG[37] (phs001790), Baron Mouse[30] (GSE84133), Baron Human[30] (GSE84133), Muraro[31] (GSE85241), Segerstolpe[32] (E-MTAB-5061), Xin[33] (GSE81608), CellBench 10X[34] (GSM3618014), CellBench CEL-Seq2[34] (GSM3618022, GSM3618023, GSM3618024), TM[6] (GSE109774), and Zheng sorted and Zheng 68K[36] (SRP073767). The PbmcBench datasets[38] are not yet uploaded to any data repository.

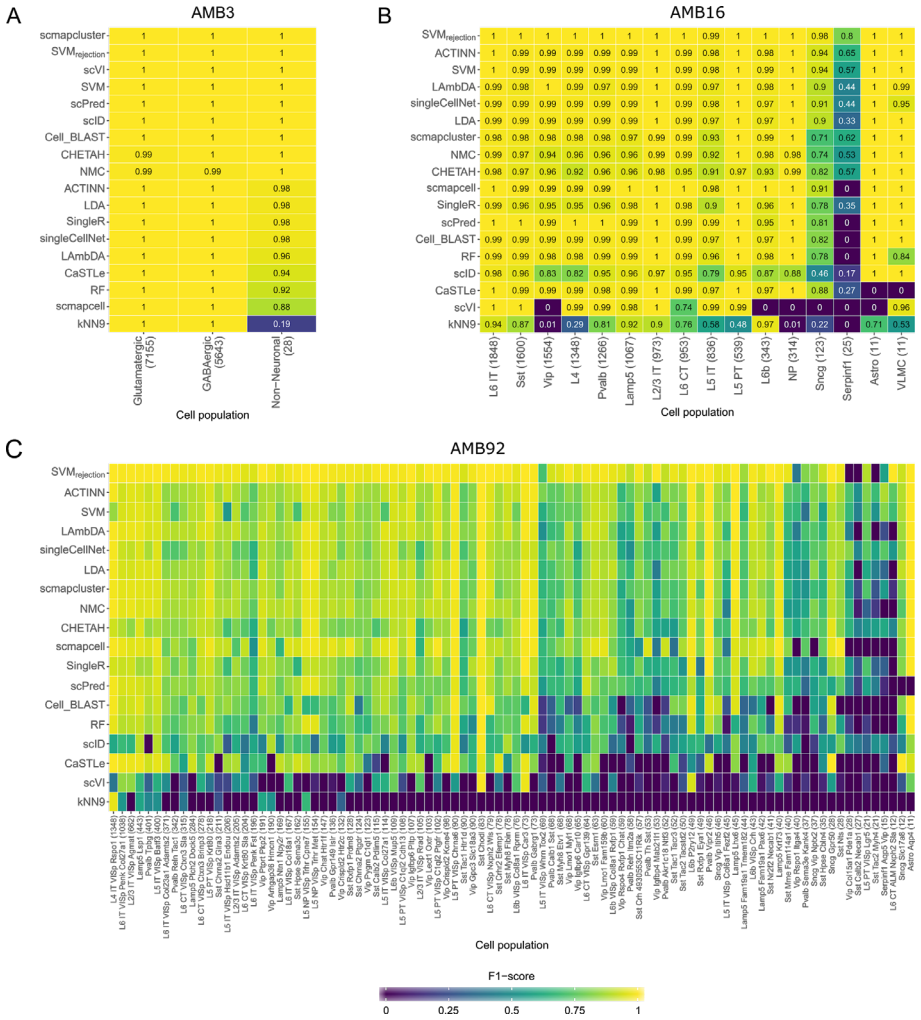
BIBLIOGRAPHY

1. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols* **13,** 599–604 (2018).
2. Plass, M. *et al.* Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science (80-.).* **360,** (2018).
3. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science (80-.).* **357,** 661–667 (2017).
4. Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M. & Reddien, P. W. Cell type transcriptome atlas for the planarian Schmidtea mediterranea. *Science (80-.).* **360,** (2018).
5. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172,** 1091–1107.e17 (2018).
6. Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562,** 367–372 (2018).
7. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566,** 496–502 (2019).
8. Henry, V. J., Bandrowski, A. E., Pepin, A. S., Gonzalez, B. J. & Desfeux, A. OMICtools: an informative directory for multi-omic data analysis. *Database (Oxford).* **2014,** (2014).
9. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput. Biol.* **14,** (2018).
10. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37,** 547–554 (2019).
11. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research* **7,** (2018).
12. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15,** 255–261 (2018).
13. Diaz-Mejia, J. J. *et al.* Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data. *F1000Research* **8,** 296 (2019).
14. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* **16,** 983–986 (2019).
15. Wagner, F. & Yanai, I. Moana: A robust and scalable cell type classification framework for single-cell RNA-Seq data. *bioRxiv* (2018). doi:10.1101/456129
16. Domanskyi, S. *et al.* Polled Digital Cell Sorter (p-DCS): Automatic identification of hematological cell types from single cell RNA-sequencing clusters. *bioRxiv* (2019). doi:10.1101/539833
17. Zhang, Z. *et al.* Scina: Semi-supervised analysis of single cells in silico. *Genes (Basel).* **10,** (2019).
18. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15,** 1053–1058 (2018).
19. Cao, Z. J., Wei, L., Lu, S., Yang, D. C. & Gao, G. Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST. *Nat. Commun.* **11,** (2020).
20. Ma, F. & Pellegrini, M. ACTINN: Automated identification of cell types in single cell RNA sequencing. *Bioinformatics* **36,** 533–538 (2020).
21. Johnson, T. S. *et al.* LAmbDA: Label ambiguous domain adaptation dataset integration reduces batch effects and improves subtype detection. *Bioinformatics* **35,** 4696–4706 (2019).
22. Kiselev, V. Y., Yiu, A. & Hemberg, M. Scmap: Projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15,** 359–362 (2018).
23. Alquicira-Hernández, J., Sathe, A., Ji, H. P., Nguyen, Q. & Powell, J. E. ScPred: Cell type prediction at single-cell resolution. *bioRxiv* (2018). doi:10.1101/369538
24. de Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T. & Holstege, F. C. P. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res.* **47,** e95 (2019).
25. Lieberman, Y., Rokach, L. & Shay, T. CaSTLe - Classification of single cells by transfer

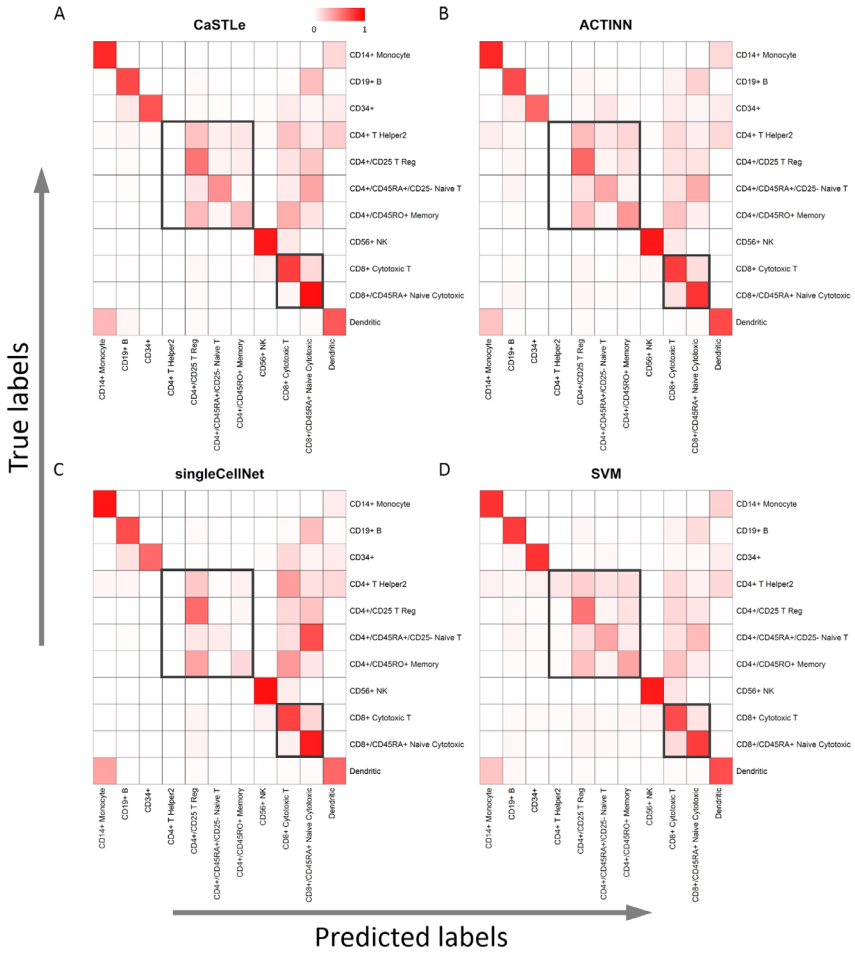
learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLoS One* **13,** (2018).
26. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20,** 163–172 (2019).
27. Boufea, K., Seth, S. & Batada, N. N. scID: Identification of transcriptionally equivalent cell populations across single cell RNA-seq data using discriminant analysis. *bioRxiv* (2018). doi:10.1101/470203
28. Tan, Y. & Cahan, P. SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. *Cell Syst.* **9,** 207–213.e2 (2019).
29. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).
30. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **3,** 346–360.e4 (2016).
31. Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst.* **3,** 385–394.e3 (2016).
32. Segerstolpe, Å. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.* **24,** 593–607 (2016).
33. Xin, Y. *et al.* RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. *Cell Metab.* **24,** 608–615 (2016).
34. Tian, L. *et al.* Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods* **16,** 479–487 (2019).
35. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563,** 72–78 (2018).
36. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8,** (2017).
37. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573,** 61–68 (2019).
38. Ding, J. *et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38,** 737–746 (2020).
39. Franzén, O., Gan, L. M. & Björkegren, J. L. M. PanglaoDB: A web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019,** (2019).
40. Zhang, X. *et al.* CellMarker: A manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* **47,** D721–D728 (2019).
41. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36,** 421–427 (2018).
42. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv* (2018).
43. Andrews, T. S. & Hemberg, M. M3Drop: Dropout-based feature selection for scRNASeq. *Bioinformatics* **35,** 2865–2867 (2019).
44. Cai, D., He, X. & Han, J. Training linear discriminant analysis in linear time. in *Proceedings - International Conference on Data Engineering* 209–217 (2008). doi:10.1109/ICDE.2008.4497429
45. Köhler, N. D., Büttner, M. & Theis, F. J. Deep learning does not outperform classical machine learning for cell-type annotation. *bioRxiv* (2019). doi:10.1101/653907
46. van den Berg, P. R., Budnik, B., Slavov, N. & Semrau, S. Dynamic post-transcriptional regulation during embryonic stem cell differentiation. *bioRxiv* (2017). doi:10.1101/123497
47. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9,** (2018).
48. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15,** (2019).
49. Finak, G. *et al.* MAST: A flexible statistical framework for assessing transcriptional changes

and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16,** (2015).
50. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36,** 411–420 (2018).
51. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **34,** 3600–3600 (2018).

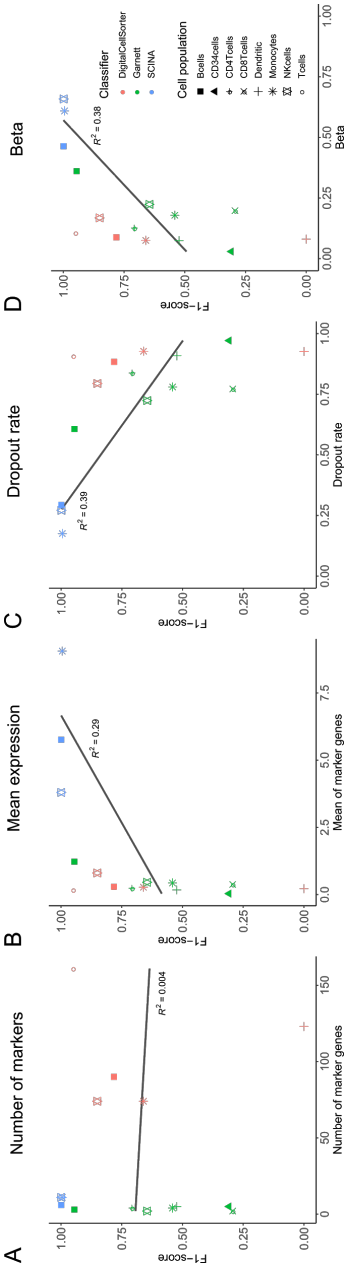
SUPPLEMENTARY MATERIALS



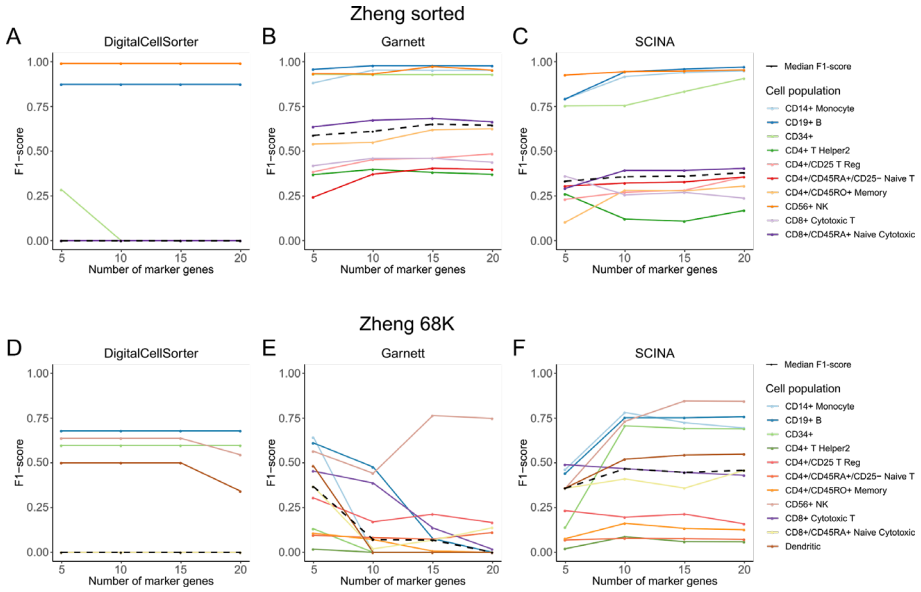
Supplementary Figure 4.1 Classification performance across different annotation levels in the Allen Mouse Brain dataset. Heatmaps show the F1-scores of each classifier for each cell population in the **(A)** AMB3, **(B)** AMB16, and **(C)** AMB92 datasets. The cell populations are sorted from left-to-right in descending order according to their size (i.e. number of cells). The size of each population is indicated between brackets. In each heatmap, the classifiers are sorted according to their mean performance across all cell populations.



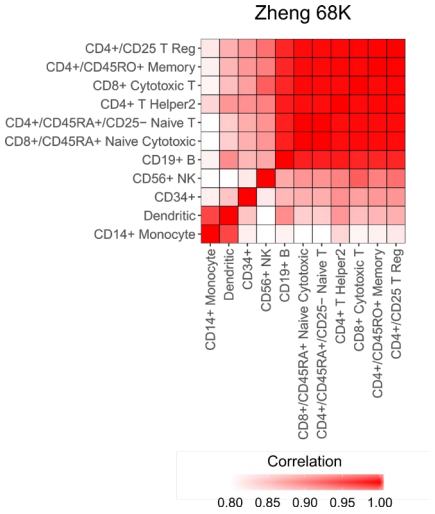
Supplementary Figure 4.2 Confusion matrices for the Zheng 68K dataset. Results of four classifiers, **(A)** *CaSTLe*, **(B)** *ACTINN*, **(C)** *singleCellNet*, and **(D)** *SVM*, are shown. Rows indicate the true labels and columns indicate the predicted labels. Each cell in the heatmap is colored according to the percentage of overlapping cells between the true and predicted cell population. Black boxes highlight the four subpopulations of CD4 and the two subpopulations of CD8 T-cells.



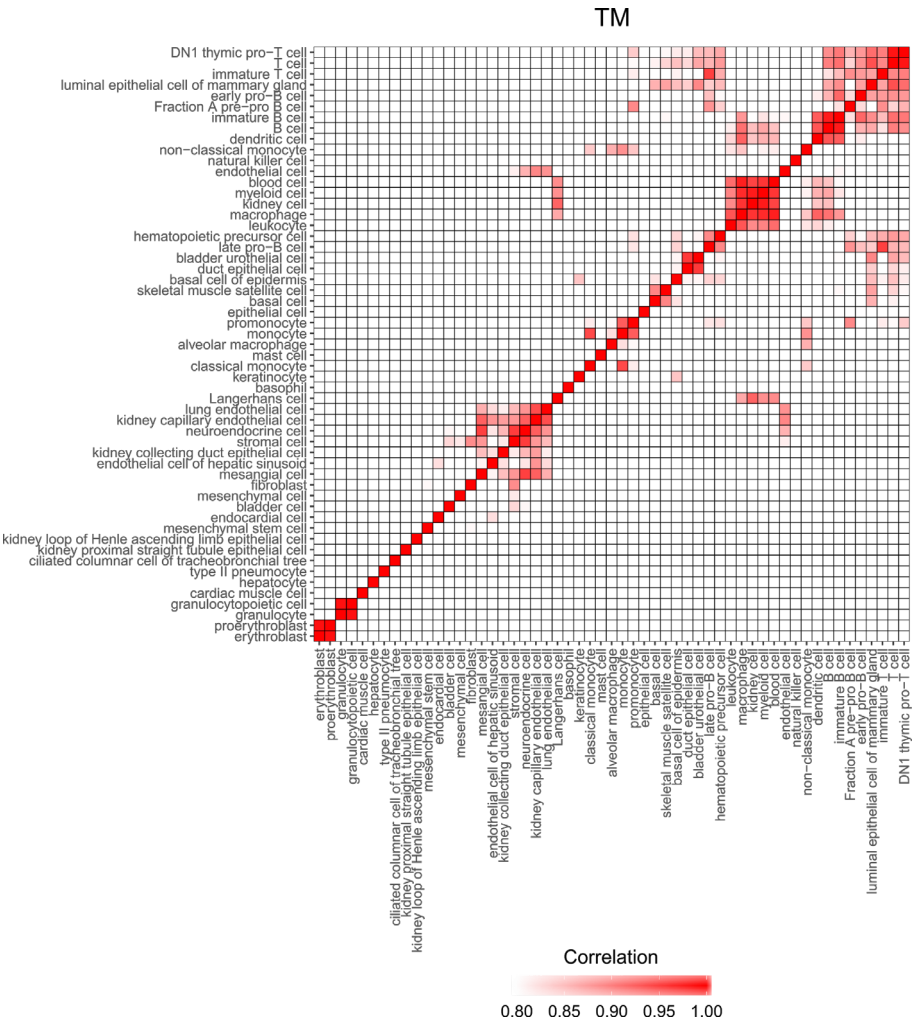
Supplementary Figure 4.3 Effect of marker-genes on the performance of the classifiers. Scatterplots compare the **(A)** number of marker-genes, **(B)** mean expression, **(C)** dropout rate, and **(D)** beta, a measure for the specificity, with the performance of the marker based classifiers. Different classifiers are indicated with different colors, different cell populations with different shapes.



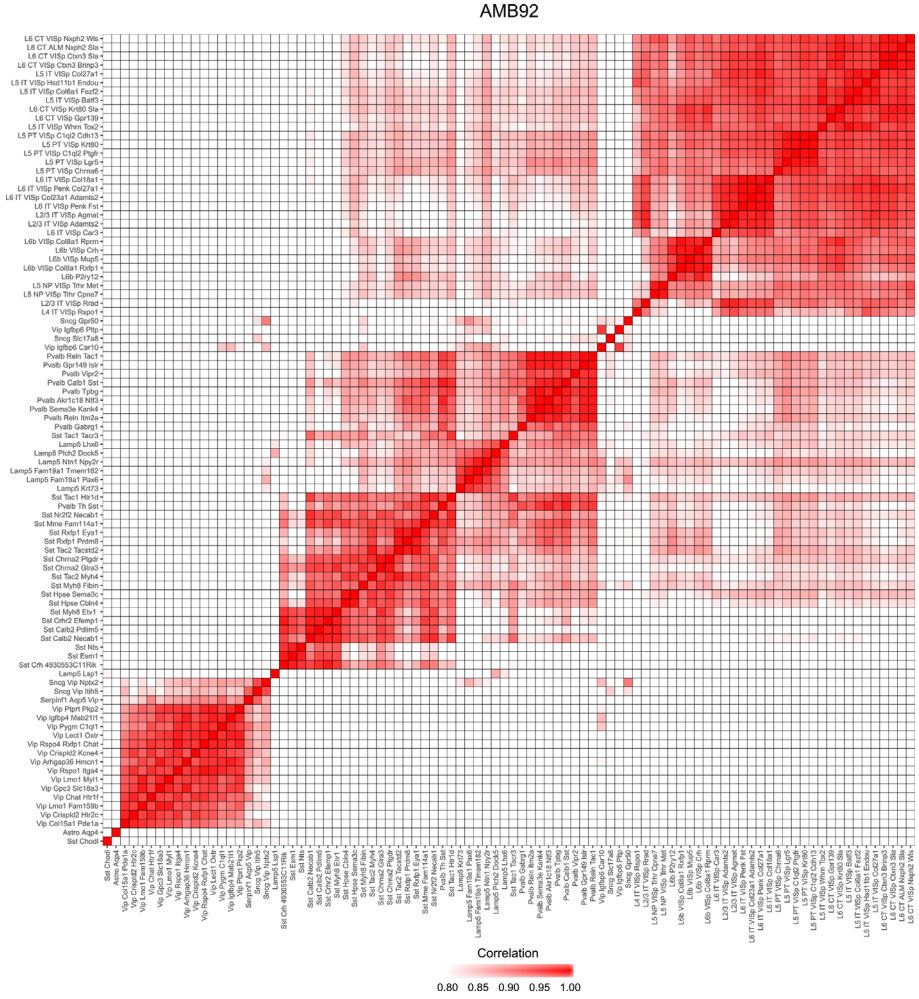
Supplementary Figure 4.4 Performance of marker-based classifiers using differentially expressed genes. Line plots show the performance of marker-based classifiers using different number of marker-genes on the **(A-C)** Zheng sorted and **(D-F)** Zheng 68K dataset. marker-genes were selected using differential expression. Different cell populations are indicated using different colors. The median F1-score of the classifier is indicated using a dashed black line.



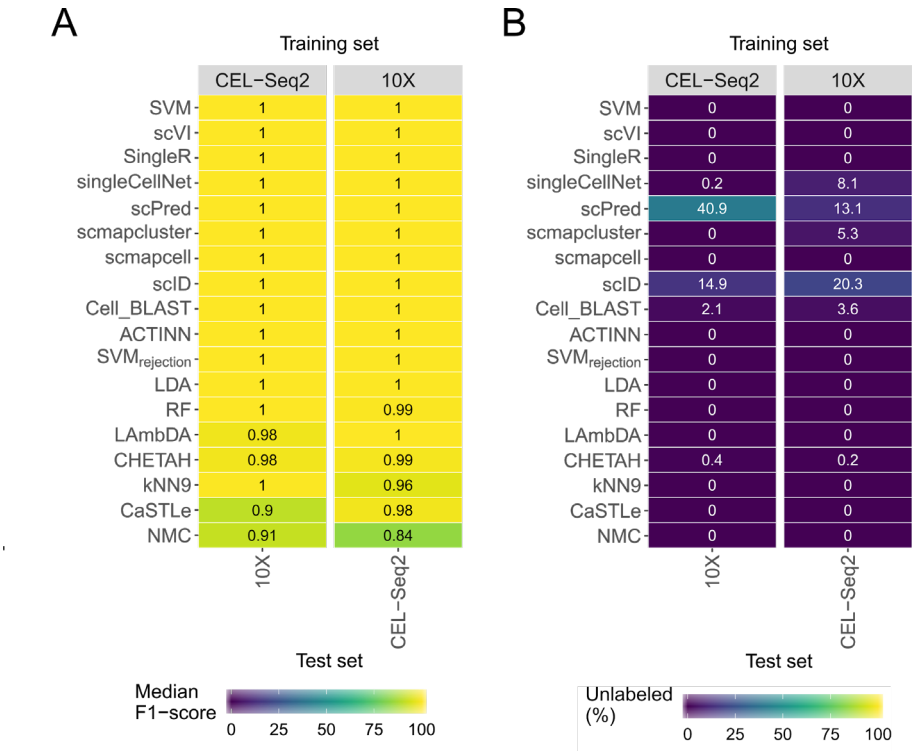
Supplementary Figure 4.5 Correlation between cell populations in the Zheng 68K dataset. Heatmap showing the pairwise Pearson correlation between the different cell populations in the Zheng 68K dataset.



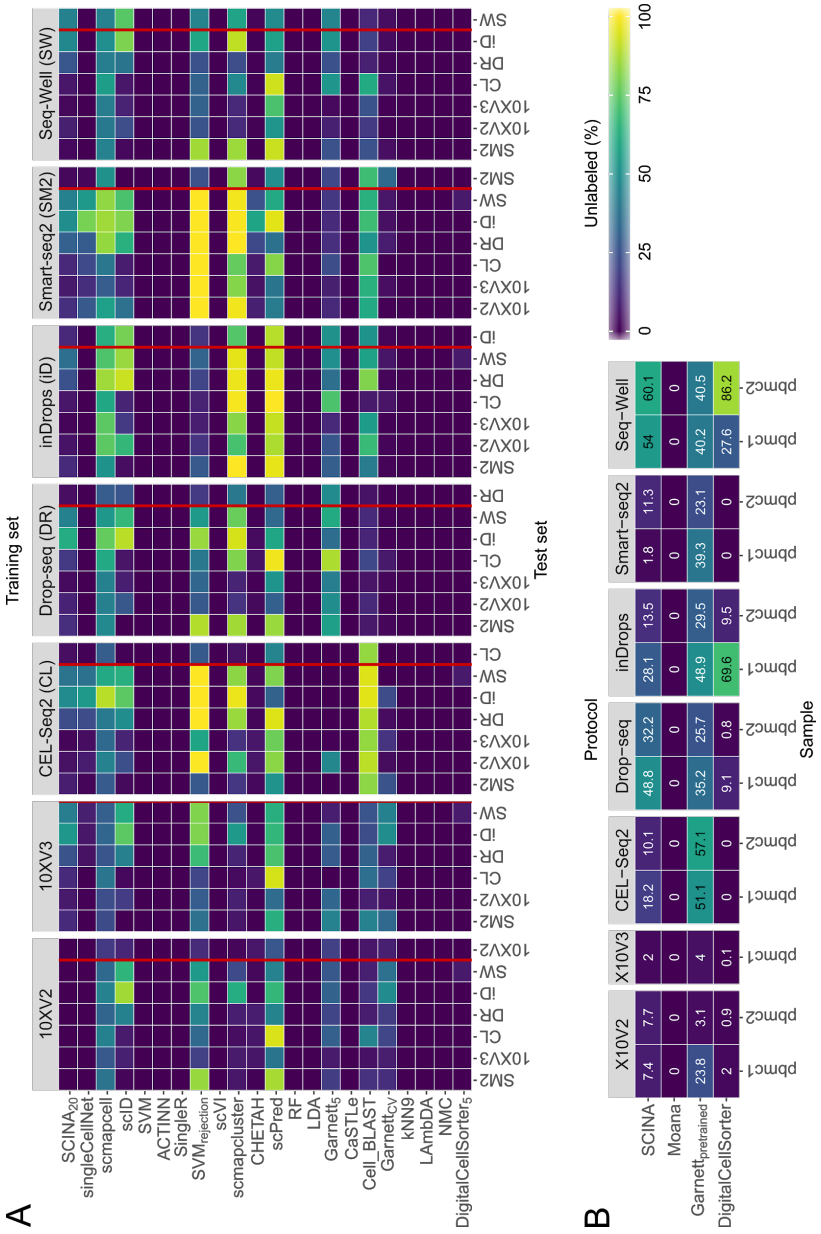
Supplementary Figure 4.6 Correlation between cell populations in the TM dataset. Heatmap showing the pairwise Pearson correlation between the different cell populations in the TM dataset.



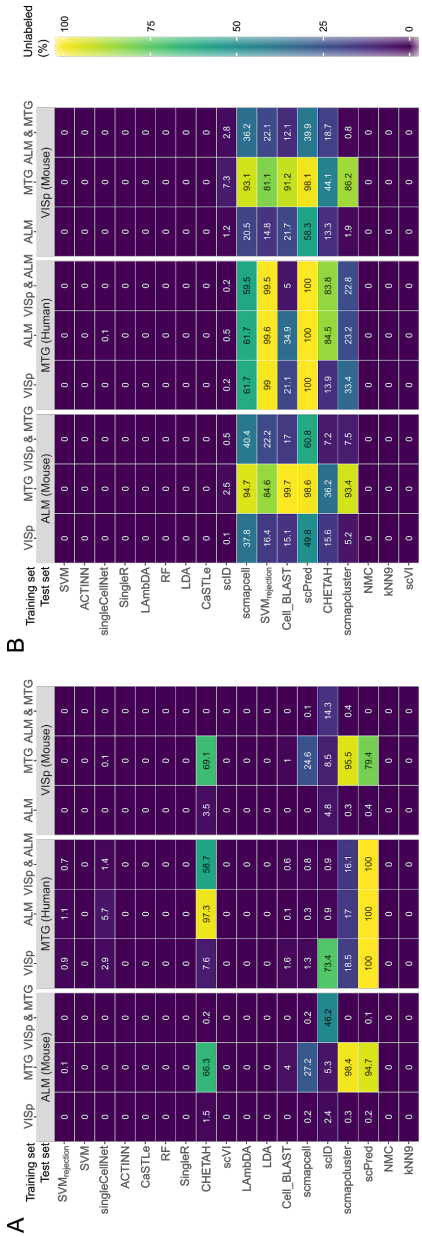
Supplementary Figure 4.7 Correlation between cell populations in the AMB92 dataset. Heatmap showing the pairwise Pearson correlation between the different cell populations in the AMB92 dataset.



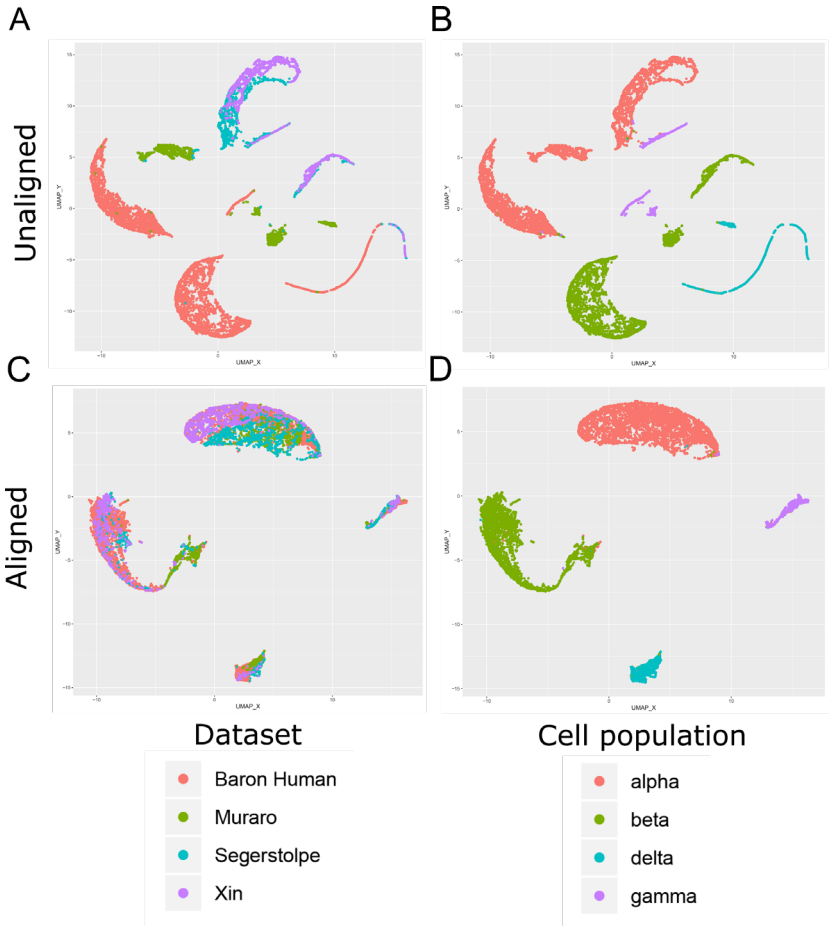
Supplementary Figure 4.8 Classification performance across the CellBench datasets. Heatmaps show the **(A)** median F1-score and **(B)** percentage of unlabeled cells across the CellBench datasets. The training set is indicated above the heatmap, the test set below. Classifiers are sorted based on their mean performance in **(A).**



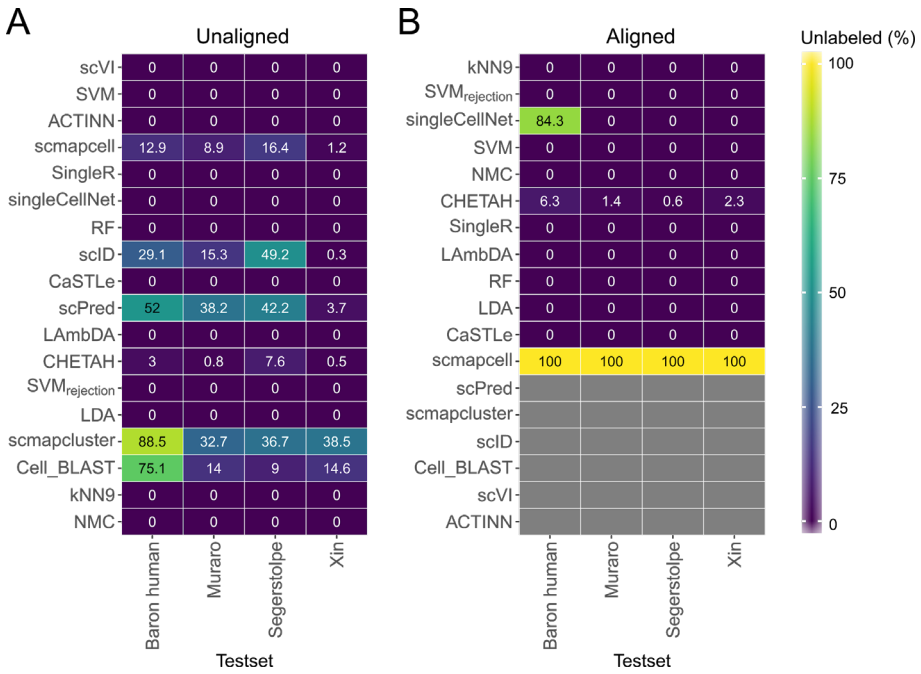
Supplementary Figure 4.9 Percentage of unlabeled cells across the PbmcBench datasets. (A) Heatmap showing the median F1-score of the supervised classifiers for all train-test pairwise combination across different protocols. The training set is indicated in the grey box on top of the heatmap, the test set is indicated using the column labels below. Results showed to the left of the red line represent the comparison between different protocol using sample pbmc1. Sample pbmc2 was used as test set then. Results showed to the right of the red line represent the comparison between different samples using the same protocol, with pbmc1 used for training and pbmc2 used for testing. For *SCINA,*, *Garnett*$_{DE}$, and *DigitalCellSorter*$_{DE}$ different numbers of marker-genes were tested. Only the best result is shown here. **(B)** Percentage of unlabeled of the prior-knowledge classifiers on both samples of the different protocols. The protocol is indicated in the grey box on top of the heatmap, the sample is indicated with the labels below. Classifiers in the heatmaps are ordered based on their mean performance in Figure 4.3.



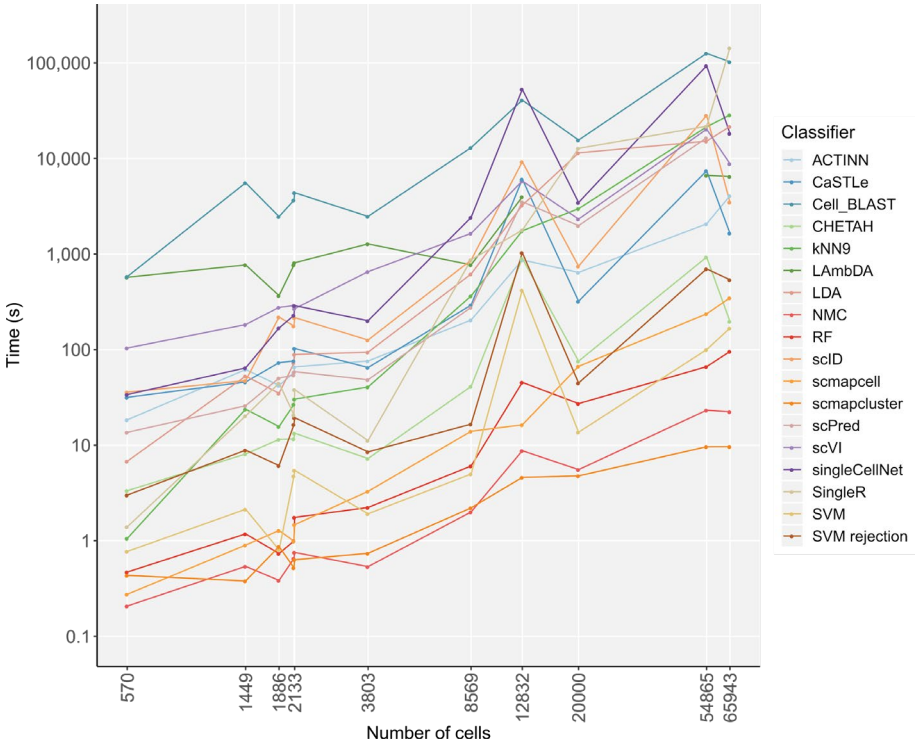
Supplementary Figure 4.10 Percentage of unlabeled across brain datasets. Heatmaps show the percentage of unlabeled of the classifiers on **(A)** major lineage annotation with three cell populations, and **(B)** deeper level of annotation with 34 cell populations. The training set(s) are indicated using the column labels on top of the heatmap. The test set is indicated in the grey box. In each heatmap the classifiers are ordered based on their mean performance in Figure 4.4.



Supplementary Figure 4.11 UMAP plots of the four pancreatic datasets used in the inter-dataset experiment. (A-B) UMAP plots before and **(C-D)** after alignment using MNN. In **(A, C)** the cells are colored by dataset and in **(B, D)** the cells are colored by cell population.



Supplementary Figure 4.12 Percentage of unlabeled cells across different pancreatic datasets. Heatmaps showing the percentage of unlabeled for each classifier for the **(A)** unaligned and **(B)** aligned datasets. The column labels indicate which of the four datasets was used as a test set, in which case the other three sets were used as training data. Grey boxes indicate that the corresponding method could not be tested on the corresponding dataset. In each heatmap, the classifiers are ordered based on their mean performance in Figure 4.5.



Supplementary Figure 4.13 Computation time across different datasets. Line plots showing the computation time of the classifiers with the number of cells in all datasets. Classifiers are indicated using different colors. Both axes in the plot are log-scaled.

CHAPTER 5

PREDICTING CELL POPULATIONS IN SINGLE CELL MASS CYTOMETRY DATA

Tamim Abdelaal

Vincent van Unen

Thomas Höllt

Frits Koning

Marcel J.T. Reinders

Ahmed Mahfouz

This chapter is published in: *Cytometry Part A* (2019) 95(7): 769-781, doi: 10.1002/cyto.a.23738.
Supplementary material is available online at:
https://onlinelibrary.wiley.com/doi/full/10.1002/cyto.a.23738

Mass cytometry by time-of-flight (CyTOF) is a valuable technology for high-dimensional analysis at the single cell level. Identification of different cell populations is an important task during the data analysis. Many clustering tools can perform this task, which is essential to identify "new" cell populations in explorative experiments. However, relying on clustering is laborious since it often involves manual annotation, which significantly limits the reproducibility of identifying cell-populations across different samples. The latter is particularly important in studies comparing different conditions, for example in cohort studies. Learning cell populations from an annotated set of cells solves these problems. However, currently available methods for automatic cell population identification are either complex, dependent on prior biological knowledge about the populations during the learning process, or can only identify canonical cell populations. We propose to use a linear discriminant analysis (LDA) classifier to automatically identify cell populations in CyTOF data. LDA outperforms two state-of-the-art algorithms on four benchmark datasets. Compared to more complex classifiers, LDA has substantial advantages with respect to the interpretable performance, reproducibility, and scalability to larger datasets with deeper annotations. We apply LDA to a dataset of ~3.5 million cells representing 57 cell populations in the Human Mucosal Immune System. LDA has high performance on abundant cell populations as well as the majority of rare cell populations, and provides accurate estimates of cell population frequencies. Further incorporating a rejection option, based on the estimated posterior probabilities, allows LDA to identify previously unknown (new) cell populations that were not encountered during training. Altogether, reproducible prediction of cell population compositions using LDA opens up possibilities to analyze large cohort studies based on CyTOF data.

5.1 INTRODUCTION

Mass Cytometry by time-of-flight (CyTOF) is a valuable tool for the field of immunology, as it allows high-resolution dissection of the immune system composition at the cellular level[1]. Advances in CyTOF technology provide the simultaneous measurement of multiple cellular protein markers (> 40), producing complex datasets which consist of millions of cells[2]. Many recent studies have shown the utility of CyTOF to identify either canonical or new cell populations while profiling the immune system. These include 1) the characterization of cell population heterogeneity for a specific cancer[3–5], 2) assigning signature cell populations when profiling a specific disease[6], and 3) monitoring the immune system response to various infections[7,8].

A key step in mass cytometry analysis is the accurate identification of cell populations in a given sample. The high number of dimensions in CyTOF data has forced researchers to depart from manual gating strategies based on two-dimensional plots because it's very labor intensive and subjective[9]. These limitations greatly impede the translational aspects of these technologies. Major efforts have been made to facilitate the analysis of CyTOF data by means of clustering (unsupervised learning) methods. These include SPADE[10], FlowSOM[11], Phenograph[4] and X-shift[12], and they are often combined with dimensionality reduction methods like PCA[13], t-SNE[14,15], and HSNE[16,17].

Clustering approaches are very instrumental in analysing high-dimensional data and identifying different cell populations in cytometry data. These populations are defined in a data-driven manner, avoiding biases arising from manual gating[18]. Thus, in explorative experiments, clustering approaches allow to identify both canonical cell populations and (new) cell populations, which is particularly useful when looking for rare populations in case-control experiments. After clustering, manual input is required to annotate the discovered cell

populations with biologically relevant labels. This can be done by visually exploring the data, either by gating the biaxial marker expression scatter plots in the case of flow cytometry (FC), by overlaying the marker expression profiles on a low-dimension representation (e.g. tSNE), or by inspecting a heatmap of the markers' expression across clusters.

Generally, this annotation process works well, especially in small explorative experiments, in which all the samples are analyzed at once. However, in cohort studies with hundreds of biological samples, the clustering analysis is usually performed per sample, or small groups of samples, as samples are collected over long time periods, or due to computational limitation in the number of cells that can be analyzed at once. Consequently, the annotation process becomes time consuming, and, more importantly, limits the reproducibility of identifying cell populations across different (batches of) samples[19]. The latter is especially pronounced when looking for deeper subtyping of cell populations rather than major populations.

These limitations are inherent to both FC and CyTOF, albeit more pronounced in the later given the higher number of dimensions and the larger number of cells being measured. In the field of FC, several supervised approaches have been proposed to automatically identify cell populations. They have been shown to match the performance of centralized manual gating based on benchmark datasets from challenges organized by the FlowCAP ("Flow Cytometry: Critical Assessment of Population Identification Methods") Consortium[20,21]. These approaches rely on learning the manual gating from a set of training samples, and transferring the learned thresholds for the gates to new test samples.

As gating is done based on two dimensional views of the data, this is not a feasible approach for CyTOF data, since the number of markers is generally around 40, resulting in $\sim 2^{40}$ of gates that need to be defined (one for every pair of markers). Moreover, manual gating generally assumes that cells of interest can be selected for by dichotomizing each marker, i.e. splitting cells on the basis of a marker being positively or negatively expressed (identified by a threshold value, i.e. the gate). However, analyses of CyTOF data have repeatedly shown that cell population composition is much more complex, showing many clusters that are described by a combination of all marker expressions[17], requiring the need for a multitude of gates that increases the complexity of gating even further.

Consequently, for CyTOF data, alternative gating approaches need to be considered. Recently, two methods have been developed: Automated Cell-type Discovery and Classification (ACDC)[22] and DeepCyTOF[23]. ACDC integrates prior biological knowledge on markers of specific cell populations, using a cell-type marker table in which each marker takes one of three states (1: positively expressed, -1: negatively expressed, 0: do not consider) for each cell population. This table is then used to guide a semi-supervised random walk classifier of canonical cell populations (i.e. cell populations with defined marker expression patterns). DeepCyTOF applies deep neural networks to learn the clustering of one sample, and uses the trained network to classify cells from different samples. Both methods achieve accurate results on a variety of datasets. However, both methods rely on sophisticated classifiers. Interestingly, neither of these methods compared their performance to simpler classifiers. Further, both methods focused mainly on classifying canonical cell populations, which is not the main focus of CyTOF studies which usually relies on the large number of markers measured for deep interrogation of cell populations.

In this work, we show that a linear discriminant analysis (LDA) classifier can accurately classify cell populations in mass cytometry datasets. Compared to previous methods, LDA presents a simpler, faster and reliable method to assign labels to cells. Moreover, using LDA instead of more complex classifiers enables the analysis of large datasets comprised of

millions of cells. To illustrate this, we tested the applicability of LDA in classifying not only the canonical cell populations but also when deeper subtyping the human mucosal immune system across multiple individuals, where the classification task becomes harder as the differences between cell populations is much smaller.

5.2 METHODS

We define a *cell* as the single measurement event in CyTOF data, $c \in \mathcal{R}^p$, where p is the number of markers on the CyTOF panel. Cells are being measured collectively from one *sample*, which is the biological specimen collected from an individual. A sample usually consists of thousands of cells, i.e. $s \in \mathcal{R}^{n_c \times p}$, where n_c is the number of cells in sample s. A CyTOF *dataset* consists of multiple samples, $d \in \mathcal{R}^{n_s \times n_c \times p}$, where n_s is the number of samples in the dataset that can comprise different groups of patients. Ultimately, we are interested in identifying cells that have a similar protein marker expression, i.e. cells that belong to the same population of cells. Note that with this definition of *cell population*, similar cells can either represent cells with the same cell type and/or state, depending on which markers are considered[24]. Usually the different cell populations are derived from clustering a large collection of cells collected from different samples using an unsupervised clustering approach.

5.2.1 DATASETS DESCRIPTION

We used four public benchmark datasets to evaluate our classifier, for which manually gated populations were available and used as ground truth reference (Supplementary Table 5.1). First, the **AML dataset** is a healthy human bone marrow mass cytometry dataset[4], consisting of 104,184 cells analyzed using 32 markers resulting in 14 cell populations defined by manual gating. Second, the **BMMC dataset** is also a healthy human bone marrow dataset[4,25], consisting of 81,747 cells analyzed with 13 markers, and 24 manually gated cell populations. Third, the **PANORAMA dataset** entails 10 replicates of mice bone marrow cells[12], analyzed using a mass cytometry panel of 39 markers and manually gated into 24 cell populations, with a total number of cells around 0.5 million. Finally, the **Multi-Center study dataset** is a collection of 16 samples drawn from a single subject[23], where the first eight samples are collected at the same time and analyzed with the same instrument, and the last eight samples are collected two months later and analyzed with a different instrument. It contains ~930,000 cells, analyzed with 26 markers, where only eight markers were used for the manual gating process[23], resulting in four canonical cell populations in addition to a fifth class representing the unlabelled cells. In addition to the benchmark datasets, we used data that we collected from patients with gastrointestinal diseases as well as controls. This **Human Mucosal Immune System mass cytometry (HMIS) dataset**[6] consists of 102 samples: 47 Peripheral Blood Mononuclear Cells (PBMC) and 55 gut tissue samples. We focused on the PBMC samples only, which are further divided into 14 control samples, 14 samples with Crohn's Disease (CD), 13 samples with Celiac Disease (CeD) and 6 samples with Refractory Celiac Disease Type II (RCDII). There are ~3.5 million cells in the 47 PBMC samples, which are measured with a panel of 28 markers. Prior to any further processing, dead cells, debris and non-gated cells were removed. Measured expressions were transformed using hyperbolic arcsin with a cofactor of 5 for all datasets.

To annotate the HMIS dataset with cell population information, we clustered all cells across all PBMC samples simultaneously using Cytosplore^{+HSNE}[26]. The motivation to choose Cytosplore^{+HSNE} is to reproduce similar cell populations to the ones defined in the original study of the HMIS dataset[6,17]. However, any other clustering method, such as FlowSOM or X-shift, could be used for this task[18]. We constructed three layers HSNE. For the top (overview)

layer, we annotated the clusters into six major immune lineages on the basis of the expression of known lineage marker: 1) CD4+ T cells, 2) CD8+ T cells, including TCRgd cells, 3) B cells, 4) CD3-CD7+ Innate Lymphocytes (ILCs), 5) Myeloid cells, and 6) Others, representing unknown cell types (Supplementary Figure 5.1). This we denoted the **HMIS-1 dataset**. Next, in order to find subtypes at a more detailed level, we explored one layer down for each of the six cell populations separately, producing six separate t-SNE maps (Supplementary Figure 5.1). For each map, we applied Gaussian Mean Shift (GMS) clustering[27], with a kernel size of 30 (default value). For each cluster, we calculated a cluster representation by taking the median expression of each marker for all individual cells annotated with that cluster. We automatically merged clusters when the correlation between cluster representatives is above 0.95. We discarded clusters containing less than 0.1% of the total number of cells (< 3500 cells). In total we ended up with 57 (clusters) cell populations (11 CD4+ T cells, 9 CD8+ T cells, 4 TCRgd cells, 11 B cells, 11 CD3-CD7+ ILCs, 6 Myeloid cells and 5 Others) for the ~3.5 million PBMC cells, which we denoted the **HMIS-2 dataset**. Cell counts per cell population and per sample are summarized in Supplementary Figure 5.2.

5.2.2 CELL POPULATION PREDICTORS

To determine cell populations in a newly measured sample, one would need to re-cluster the new sample with all previous samples. Besides being a tedious task, cells from the new sample will influence the clustering and by that change the previously identified cell populations, affecting reproducibility. Therefore, we learn the different cell populations from a training set with annotated cells. The cell populations in the new sample can then simply be predicted by this learned cell- populations predictor.

LDA. We propose to use a (simple) Linear Discriminant Analysis (LDA) classifier to predict cell populations in CyTOF data. To produce a cell population prediction for new cell x, LDA assign x to cell population class c_i for which the posterior probability of x being part of c_i is maximum, across all cell populations.

Assign x to $\arg\max_{\forall c_i} p(x|c_i)P(c_i)$

where $p(x|c_i) = \frac{1}{(2\pi)^{k/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}$, $\quad \Sigma_i = \Sigma \quad \forall c_i$

$P(c_i)$ is the prior probability of cell population class c_i, which is equal to the number of cells in cell population i divided by the total number of cells in the dataset, k is the number of features (protein markers in case of CyTOF), μ_i is the k-dimensional mean vector of cell population class c_i, Σ_i is the $k \times k$ covariance matrix of cell population class c_i.

k-NN. Further, to check whether the performance of a non-linear classifier would outperform the linear LDA classifier, we tested the performance of a k-NN classifier (with Euclidean distance and k = 50 neighbours). We adopted an editing approach when training the k-NN classifier to reduce the training set size, and consequently keep testing times reasonable. The editing is done according to the following pseudo code. We start by creating a training set *(Tr)*, by sampling 50,000 cells uniformly and without replacement from all samples in the original training data *(OrgTr)*. Next, we create a test set *(Te)*, by sampling another 50,000 cells uniformly and without replacement from *OrgTr*. The k-NN classifier is then trained using *Tr* and used to make cell population predictions for *Te*. All correctly predicted cells from *Te* are ignored while the misclassified cells are added to *Tr*. We iterate these steps until there are no cells left within *OrgTr*, i.e. we have processed all cells. The final version of *Tr* contains

much less cells than the original *OrgTr*, but will encompass the necessary representative cells from each cell population class to achieve a similar k-NN performance.

```
Input: Training_Data used to train the k-NN classifier

Output: reduced version of the Training_Data representative for the input data

BEGIN

    Temp_Training ⟵ random 50,000 cells from Training_Data

    while (not all Training_Data is covered)

        Temp_Testing ⟵ another random 50,000 cells from Training_Data

        Apply prediction on Temp_Testing and add misclassified cells to
    Temp_Training

        Temp_Training ⟵ Temp_Training + Misclassified from Temp_Testing

    end while

    Final_Training ⟵ Temp_Training

END
```

NMC. Also, we tested whether an even simpler classifier, than LDA, would be sufficient to accurately identify cell populations. We tested the Nearest Median Classifier (NMC) which assigns each cell to the nearest median (median expression across all cells for a cell population) using $(1-\rho)$ as distance, with ρ being the Pearson correlation between the two expression vectors[28].

5.2.3 PERFORMANCE METRICS

To evaluate the quality of the classification, we used four metrics:

i) The classification accuracy (fraction of correctly identified cell).

ii) The F1-score (harmonic mean of the precision and recall) for which we report the median value across all cell populations. When comparing to DeepCyTOF[23], we use the weighted average of F1-scores per cell population size, to produce a fair comparison.

$$Weighted\ F1\ score = \sum_i \frac{n_i}{N} F_i$$

where n_i is the number of cells in population i, N is the total number of cells in the dataset, and F_i is the F1-score for cell population i.

iii) The maximum difference in population frequencies, defined as $\Delta f = max_i|f_i - \hat{f}_i|$, where f_i and $\hat{f}_i$ represents the true and the predicted percentage cell frequencies for the *i*-th cell population, respectively.

iv) The Root of Sum Squared Error (RSSE) per sample and per cell population, defined as $RSSE = \sqrt{\sum_{i=1}^{n}(f_i - \hat{f}_i)^2}$. In case of measuring the error per sample, f_i and $\hat{f}_i$ represents the true and the predicted percentage cell frequencies, respectively, for the *i*-th cell population per sample, and $n = n_t$ (total number of cell populations). In case of measuring the error per cell population, f_i and $\hat{f}_i$ represents the true and the predicted percentage cell frequencies, respectively, for a certain cell population in the *i*-th sample, and $n = n_s$ (total number of samples).

5.2.4 PERFORMANCE ESTIMATION

The performance of a classifier is evaluated using three different cross-validation setups:

i) CV-Cells: 5-fold cross validation applied over all the cells.

ii) CV-Samples: A leave-sample-out cross validation over all the samples, regardless of the number of cells within each sample. The classifier is trained using the cells of the samples in the training set, then the cell population prediction is done per left-out sample.

iii) Conservative CV-Samples: Similar to CV-Samples, but with the main difference that the ground-truth reference labels, acquired by clustering, are not used for training. Instead, for each set of training samples the data is re-clustered, resulting in new cell populations. These new cell populations are then used to train the classifier, which is subsequently used to predict the labels of the cells of the left-out sample. Since the labels of the training set and the ground-truth are now different, we matched the cluster labels by calculating their pairwise correlation (Pearson's r) using the median marker expression of each cluster. Each training cluster is matched to the ground-truth cluster with which the correlation is maximum.

For the AML and the BMMC datasets, we evaluated the performance using the *CV-Cells* setup only, since no sample information is provided. For the PANORAMA and Multi-Center datasets, we used both the *CV-Cells* and *CV-Samples* setups, since we have the sample information. Considering the number of samples in each dataset, we used a 5-fold *CV-Samples* for the PANORAMA dataset and a 4-fold *CV-Samples* for the Multi-Center dataset. For the HMIS-1 and HMIS-2 datasets, we used all three cross validation setups, using a 3-fold *CV-Samples* and *Conservative CV-Samples*.

5.2.5 REJECTION OPTION

To be able to detect new cell populations, we decided to include a rejection option for LDA by defining a minimum threshold for the posterior probability of the assigned cell populations. Thus, a cell is labelled as 'unknown' whenever the posterior probability is less than a predefined threshold set.

$$\text{Assign } x \text{ to } \begin{cases} \arg\max\limits_{\forall c_i} p(x|c_i)P(c_i), & \max\limits_{\forall c_i} \frac{p(x|c_i)P(c_i)}{p(x)} > threshold \\ unknown, & otherwise \end{cases}$$

5.2.6 FEATURE SELECTION

To avoid overfitting, we explored the need to reduce the number of markers (i.e. features) by applying feature selection on the training data. First, we applied a 5-fold *CV-Cells* and used the classification performance for every individual marker on the training data to rank all

markers in a descending order. Next, we applied another 5-fold *CV-Cells* on the training data and trained as many classifiers as there are markers. The first classifier is based on the top marker only, the second one on the two top ranked markers, etc. Then we select the classifier which generates the best cross validation performance over the training set. This classifier is subsequently tested on the test set and the performance is reported.

5.3 RESULTS

5.3.1 LDA OUTPERFORMS COMPLEX CLASSIFICATION APPROACHES

To evaluate the performance of the LDA classifier, we compared LDA with two recent state-of-the-art methods for classifying CyTOF data, ACDC[22] and DeepCyTOF[23]. We used the AML, BMMC and PANORAMA datasets (used by ACDC) and the Multi-Center dataset (the only available dataset used by DeepCyTOF). We compared the performance of LDA with our reproduced values, and the reported values in these two studies (Table 5.1). ACDC was applied only for the AML and BMMC datasets, for which the cell-type marker table was provided.

Since there was no sample information available for the AML and BMMC datasets, we evaluated the performance of the LDA classifier on both datasets using the *CV-Cells* setup only, and we are unable to run DeepCyTOF on those datasets. For the AML dataset, LDA achieved comparable performance in terms of accuracy and median F1-score to ACDC. For the BMMC dataset, we applied the LDA classifier to classify all 24 cell populations, resulting in ~ 96% accuracy and 0.85 median F1-score. To have a fair comparison with ACDC, we also considered four populations as unknown[22] then classified only 20 cell populations. In both cases, LDA outperformed ACDC, specially based on the median F1-score. Similar conclusions can be observed when looking at the detailed performance per cell population, showing comparable performance for AML dataset (Figure 5.1A), and performance improvement for small populations in BMMC dataset (smallest 10 populations in Figure 5.1B).

Table 5.1 Performance summary of LDA versus ACDC, DeepCyTOF and NMC

	LDA *CV-Cells*	LDA *CV-Samples*	ACDC[a]	Deep-CyTOF[b]	NMC
			Accuracy		
AML	98.13±0.09	n.a.	98.33±0.02 98.30±0.04[c]	n.a.	97.34±0.08
BMMC	95.82±0.10 95.61±0.16[d]	n.a.	93.20±0.70 92.90±0.50[c]	n.a.	85.83±0.21
PANORAMA	97.16±0.07 97.70±0.03[d]	97.22±0.31 97.67±0.29[d]	n.r.	n.a.	94.72±0.54
Multi-Center	98.51±0.04	98.44±1.66 98.82±1.73[f]	n.a.	n.r.	98.24±1.86
			Median F1-score		
AML	0.95	n.a.	0.94 0.93[c]	n.a.	0.93
BMMC	0.85 0.85[d]	n.a.	0.69 0.60[c]	n.a.	0.62
PANORAMA	0.93 0.95[d]	0.93 0.95[d]	0.88[c]	0.59±0.01[e]	0.89
Multi-Center[b]	0.99	0.99 0.98[f]	n.a.	0.97±0.01[e] 0.93[c]	0.98

n.a. = not available, n.r. = not reported.

[a]The ACDC performance values represent the training performance, [b]Weighted F1-score, [c]Reported values in the original study, [d]Classes considered unknown, similar to ACDC, [e]Mean ± std of 10 different runs, [f]Only one sample is training (Sample 2), similar to DeepCyTOF.

On the PANORAMA dataset, we tested the LDA classifier to classify all 24 populations using both the *CV-Cells* and *CV-Samples* setups. In addition, we tested the performance of LDA on 22 populations only to have a fair comparison with ACDC[22]. In both cases LDA produces relatively high accuracy and median F1-score, and outperformed ACDC and DeepCyTOF in terms of the median F1-score (no accuracy reported by ACDC). Across all cell populations, LDA has a large F1-score improvement compared to DeepCyTOF (Figure 5.1C).

For the Multi-Center dataset, we applied *CV-Cells* and *CV-Samples* yielding an accuracy of ~98% and weighted F1-score of 0.99 for both setups. To have a fair comparison with DeepCyTOF, we only used sample no. 2 for training and tested the performance of LDA on the other 15 samples. Following DeepCyTOF, the 'unlabelled' class was excluded from the training data and during testing any prediction with probability less than 0.4 was considered 'unlabelled'. Next, the 'unlabelled' class was excluded while calculating the cell population precisions. Overall, LDA achieved comparable performance to DeepCyTOF on the Multi-Center dataset (Table 5.1, Figure 5.1D), using a denoising encoder and excluding the additional calibration step[23]. Also, DeepCyTOF suffers from lack of reproducibility, producing different results in each run, which is not the case for LDA (Figure 5.1C-D). Further, similar to DeepCyTOF, LDA has better performance on samples from the same batch as the training sample compared to samples from a different batch (Supplementary Figure 5.3).

5.3.2 LDA ACCURATELY CLASSIFIES IMMUNE CELLS IN A LARGER DATASET WITH DEEPER ANNOTATION OF CELL SUBTYPES

To test our hypothesis that LDA can achieve acceptable performance on large datasets and with more detailed cell subtyping, we applied LDA to the HMIS dataset comprised of ~3.5 million cells. The HMIS data was clustered at two levels of detail (see Methods) resulting in two different annotations for the HMIS data set: HMIS-1, representing six major lineages, and HMIS-2 containing 57 cell populations. For both annotations, we applied all three cross validation setups, *CV-Cells*, *CV-Samples* and *Conservative CV-Samples* (Table 5.2).

We first tested the LDA performance on HMIS-1, hence only classifying the canonical cell populations. LDA achieved an accuracy > 99% and a median F1-score > 0.98 for both *CV-Cells* and *CV-Samples*. Next, we applied LDA to HMIS-2, which implied classifying cells into 57 different cell populations including abundant and rare cell populations. As expected, LDA had a lower performance on HMIS-2 compared to HMIS-1 using both *CV-Cells* and *CV-Samples*, with an accuracy ~86% and a median F1-score ~0.80 (Table 5.2). The confusion matrix shows that the performance drop between HMIS-1 and HMIS-2 is mainly caused by misclassifications within the same major lineages (Supplementary Figure 5.4A). We further investigated the LDA performance across different sample types (Control, CeD, RCDII and CD) in the HMIS dataset. Figure 5.2A shows that LDA has the highest accuracy for the control samples, while the lowest accuracy is for the RCDII samples.

To better mimic a realistic scenario and avoid any leakage of information from the testing samples by considering all samples when pre-clustering cells to determine the ground truth labels, we used a *Conservative CV-Samples* setup to evaluate the LDA classifier (see Methods). For the HMIS-1 dataset representing the major lineages, the performance of LDA in the *Conservative CV-Samples* was comparable to the other setups (*CV-Cells* and *CV-Samples*), Table 5.2. The performance of the LDA classifier dropped when considering the *Conservative CV-Samples* setup on HMIS-2 that contains a multitude of cell populations. However, the lower performance can be explained by miss-matching clusters between the training set and the ground-truth, which introduces classification errors. For example, cluster 'CD4 T 11' is never predicted by the classifier, which means all cells falling within this cluster

will be misclassified (Supplementary Figure 5.4B). This is because in all 3-folds, no training cluster matches to this ground-truth cluster 'CD4 T 11' (Supplementary Figure 5.5). Whereas in case of HMIS-1, with only six dissimilar clusters, the clusters map works perfectly, resulting in high performance (Supplementary Figure 5.6).

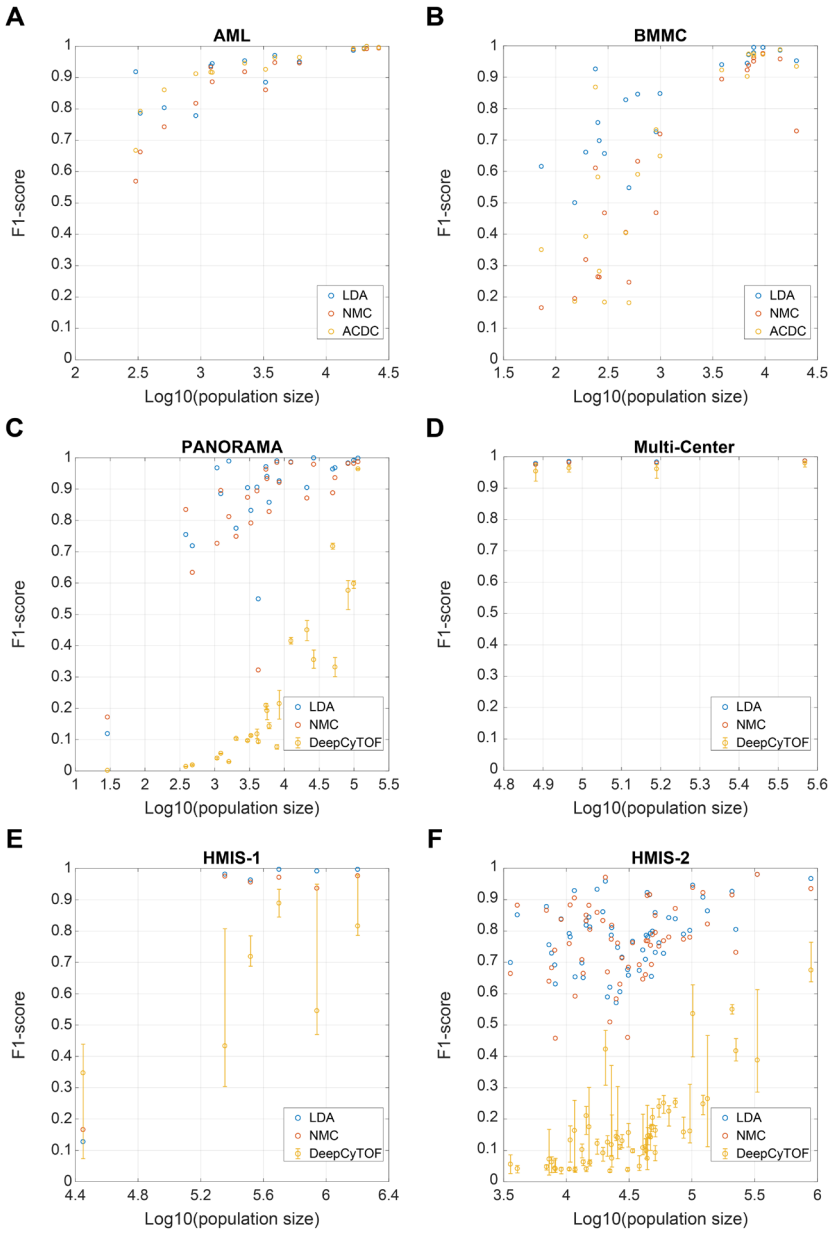


Figure 5.1 Classifiers performance comparison. Scatter plots of the F1-score vs. the population size for (**A**) AML, and (**B**) BMMC, between LDA, NMC, and ACDC. Scatter plots of the F1-score versus the population size for (**C**) PANORAMA, (**D**) Multi-Center, (**E**) HMIS-1, and (**F**) HMIS-2, between LDA, NMC, and DeepCyTOF. Error bars for DeepCyTOF shows the maximum and the minimum performance across 10 different runs.

We compared the performance of LDA on the HMIS dataset with DeepCyTOF (Table 5.2, Figure 5.1E-F). For both HMIS-1 and HMIS-2 datasets, LDA outperforms DeepCyTOF, which particularly shows a poor performance for the deeply annotated HMIS-2 dataset. These results show that LDA is robust and scalable to large datasets with deep subtyping of cell populations.

Table 5.2 Performance summary of LDA, DeepCyTOF, NMC, and k-NN on the HMIS dataset

	HMIS-1		HMIS-2	
	Accuracy	Median F1-score	Accuracy	Median F1-score
LDA *CV-Cells*	99.38±0.01	0.99	87.19±0.05	0.81
LDA *CV-Samples*	99.02±2.26	0.99 (0.98[a])	86.11±3.86	0.79 (0.87[a])
LDA *Conservative CV-Samples*	98.91±1.87	0.99	78.69±8.65	0.62
DeepCyTOF[a]	n.a.	0.72±0.06[b]	n.a.	0.36±0.02[b]
NMC	96.42±3.19	0.96	83.34±4.11	0.77
k-NN *CV-Samples*	n.a	n.a.	87.73±4.09	0.81
k-NN *CV-Samples* with feature selection	n.a.	n.a.	86.33±3.17	0.79

n.a. = not available.
[a]Weighted F1-score.
[b]Mean ± std of 10 different runs.

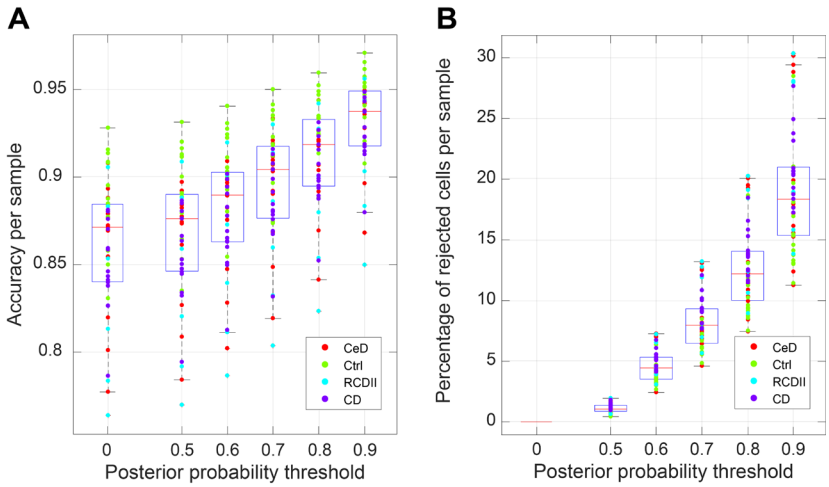


Figure 5.2 LDA accuracy and rejection size per sample. (**A**) boxplot of the LDA accuracy distribution per sample, while using a rejection threshold (0 = no rejection). (**B**) Boxplot of the rejection percentage per sample while using a rejection threshold (0 means no rejection). Each dot represents a sample colored according to the sample type (CeD: celiac disease; Ctrl: control; RCDII: refractory celiac disease type II; CD: Crohn's disease).

5.3.3 LDA OUTPERFORMS SIMPLER CLASSIFIERS

In order to explore to what extent a simple classifier can achieve high performance on identifying cell populations, we tested the NMC on all datasets. Our results show that the NMC has a comparable performance with the LDA on the Multi-Center and HMIS-1 datasets (Table 5.1 and 5.2, Figure 5.1D-E). However, LDA outperforms NMC on the AML, BMMC and PANORAMA datasets (Table 5.1, Figure 5.1A-C). Similar to ACDC, NMC suffers from large performance drop for the 10 smallest populations in the BMMC dataset (Figure 5.1B). Also, LDA outperforms NMC on the deeply annotated HMIS-2 dataset, showing performance improvement for the majority of the 57 cell populations (Table 5.2, Figure 5.1F). These results show that a simpler classifier such as NMC can predict major lineages but are not sufficient to classify deeper annotated CyTOF datasets containing smaller (rare) cell populations.

5.3.4 LDA ACCURATELY ESTIMATES CELL POPULATION FREQUENCIES

One of the main aims of CyTOF studies is to estimate the frequencies of different cell populations in a given sample. We evaluated the LDA prediction performance in terms of predicted population frequencies, by calculating the maximum difference in population frequencies, Δf, for each dataset (see Methods). LDA produced comparable population frequencies to the manually gated populations, with Pearson R correlation >0.97, between the true and predicted population frequencies for all datasets (Figure 5.3). We observed that some cell populations are harder to predict, including: 1) small populations, such as MPP in the BMMC dataset, and HSC and CLP in the PANORAMA dataset; and 2) populations that have similar cell populations in the dataset, such as 'B-cell Frac A-C (pro-B cells)' in the PANORAMA dataset, where ~41% of the cells were misclassified into the similar B cell subtypes (IgD- IgMpos B cells, IgDpos IgMpos B cells, and IgM- IgD- B cells), having a correlation of 0.86, 0.70 and 0.90 with 'B-cell Frac A-C (pro-B cells)', respectively. Overall, The maximum difference in population frequency (Δf) was 0.40%, 0.65%, 0.64% and 0.83% for the AML, BMMC, PANORAMA and the Multi-Center datasets, respectively.

For the HMIS-1 dataset, LDA has Δf of 0.59% across the six major cell populations. Interestingly, despite the drop in the accuracy of predicting cell labels on HMIS-2 compared to HMIS-1, the population frequencies are not significantly affected. The maximum difference of population frequencies in HMIS-2 was 0.46% among all 57 cell populations (Figure 5.3F). This small Δf shows that LDA produces accurate performance with respect to the ground-truth reference, even at a detailed annotation level.

We investigated the population differences per sample and per cell population using the *CV-Samples* setup in the HMIS-2 dataset, by calculating the average squared differences between the estimated and true frequencies (RSSE, see Methods). We obtained small RSSE values with a maximum of 0.074 (sample no. 10) and 0.082 ('Myeloid 10' population) across different samples and different cell populations, respectively (Supplementary Figure 5.7). For sample no. 10, the maximum absolute population difference was 5.17% for 'Myeloid 3' cell population. For 'Myeloid 10' cluster, the maximum absolute difference 5.12% across all cells.

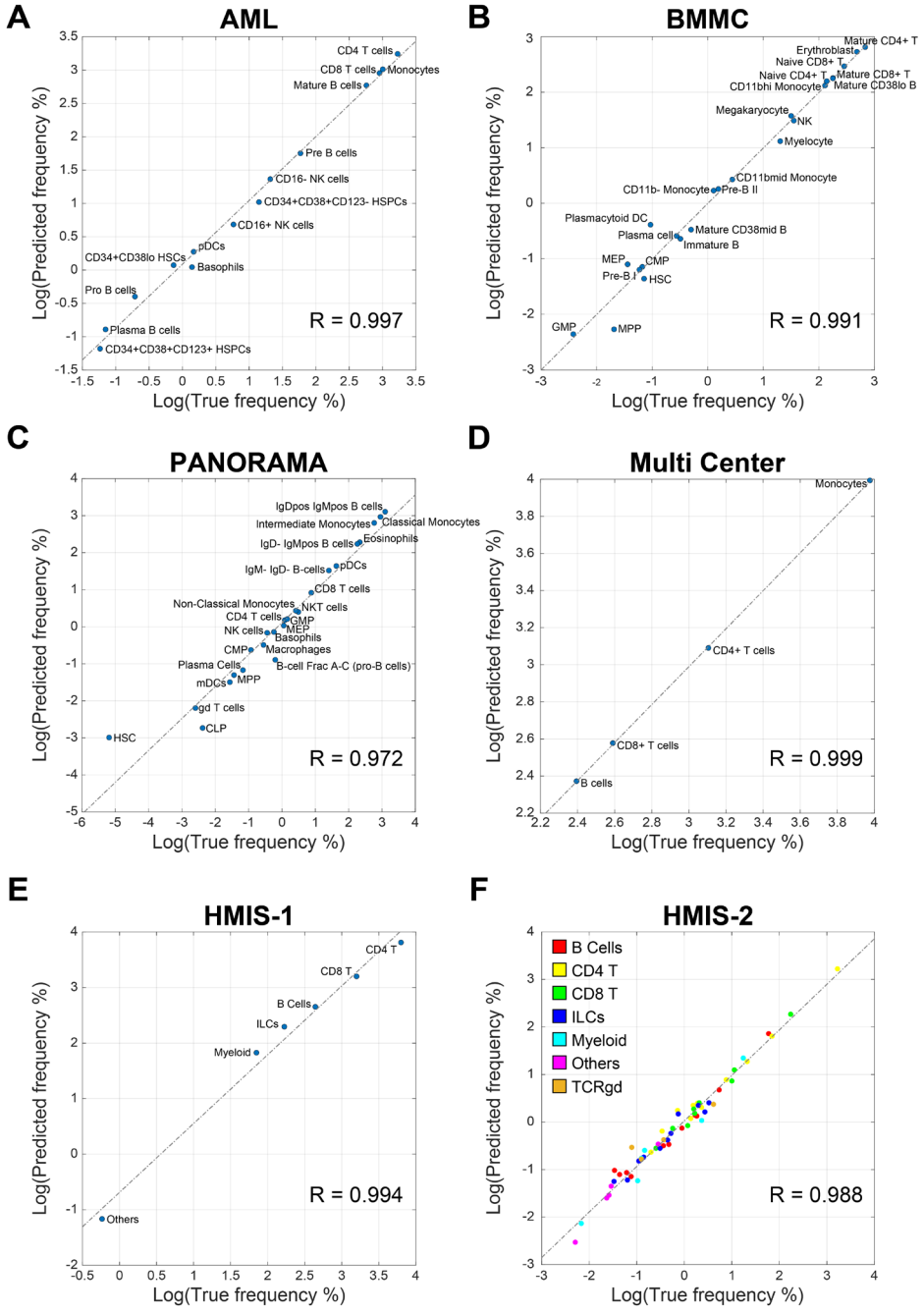


Figure 5.3 Scatter plots between true and predicted population frequencies. (**A**) AML, (**B**) BMMC, (**C**) PANORAMA, (**D**) Multi-Center, (**E**) HMIS-1, and (**F**) HMIS-2. In each plot, the dashed line shows the least-squares fit error line, and the *R* value represents Pearson correlation coefficient between true and predicted frequencies.

5.3.5 LDA PERFORMS ON HIGHLY ABUNDANT AS WELL AS RARE CELL POPULATIONS

To evaluate the performance of LDA for abundant and rare cell populations, we investigated the F1-score per cell population versus the population size. Figure 5.1F and Supplementary Figure 5.8A, show the F1-score for all 57 cell populations in the HMIS-2 dataset obtained using the *CV-Samples*. Remarkably, LDA performs well for large cell populations, as well as the majority of the small cell populations, with a median F1-score of 0.7915 for populations that contain less than 0.5% of the total cells.

For the *Conservative CV-Samples* setup, the LDA performance is still high for large cell populations, but the F1-score drops for small populations reinforcing that the drop in performance of the *Conservative CV-Samples* is driven by the limitations with the cluster matching rather than the performance of the LDA (Supplementary Figure 5.8B). For populations that contain less than 0.5% of the total cells, the median F1-score is 0.4753. Similar patterns were observed for the other four datasets (Figure 5.1A-D).

5.3.6 LDA AS A PROBABILISTIC CLASSIFIER DIRECTLY ALLOWS THE DETECTION OF UNSEEN CELL POPULATIONS

A major advantage of clustering and visual analytics over classification approaches is the ability to identify novel unknown cell populations. Here, we show that LDA as a probabilistic classifier can be used to flag unknown cells that do not match any of the training cell populations. We incorporated a rejection option to allow the classification of a cell as 'unknown' when the posterior probability of the classification of any cell is low. Figure 5.2A shows the classification accuracy across samples from the HMIS-2 dataset, after excluding unknown cells for which the posterior probability is lower than a certain threshold. As expected, setting a threshold on the posterior probability resulted in more accurate predictions. For example, setting a threshold at 0.7 resulted in an accuracy of 89.54 ± 3.25% (compared to 86.11 ± 3.86% without any thresholds), while assigning ~8% of cells per sample as unknown. The performance improvement per population shows very little variation among all the 57 cell populations (Supplementary Figure 5.9A). The difference in F1-scores, between having no rejection and applying a threshold of 0.7, is 0.04±0.02. This result shows that the rejection is not related to the overall population size, which can also be observed when calculating the rejected percentage of cells per cell population (Supplementary Figure 5.9B).

Further, we observed a reverse pattern between the accuracy of cell classification and the percentage of cells classified as unknown per sample (Figure 5.2A-B). For instance, LDA has the highest accuracy on classifying cells from the control samples and hence control samples are less likely to entail rejected (unknown) cells. On the other hand, the accuracy is the lowest on RCDII samples which also have the highest rejection percentages. Figure 5.2 further shows that both the accuracy and the rejection size increase with increasing the minimum threshold of the posterior probability.

5.3.7 REJECTION OPTION TARGETS RARE SAMPLE-SPECIFIC CELL POPULATIONS

Next, we investigated the effect of the rejection option on rare and abundant cell populations. In the HMIS-2 dataset, the population frequencies of the 57 cell populations varied from 25.2% to 0.1% of the total number of cells in the HMIS-2 dataset (Figure 5.4A). Further, we observed a variable distribution of cell populations across different sample types (control, CeD, RCDII and CD), Figure 5.4B. Although the majority of cell populations were evenly

distributed over all samples, some were disease-specific, especially the rare cell populations. Using a rejection threshold of 0.7, we calculated the rejection ratio per cell population per sample (Figure 5.4D) as the number of cells assigned as 'unknown' of one cell population in one sample, divided by the total number of cells of that cell population in all samples. We compared these rejection ratios with the cell population frequencies over the samples (Figure 5.4C) where a value close to 100% means that the cell population is specific to only one sample. We observed a strong correlation between the cell population rejection ratios and the frequencies over the samples (Figure 5.4E). For example, the majority of 'Others 2' (83.87%) comes from one CeD sample, within which 'Others 2' is prominently present (7.44% of the cells in this sample belong to 'Others 2' Supplementary Figure 5.2). The classifier rejects ~15% of these cells, representing a ~12% rejection ratio of the total number of 'Others 2' cells. This is a relatively high rejection percentage compared to other cell populations (Figure 5.4E). The main reason why there is a large rejection ratio for these cells, is because these cells are mainly present in one sample. When this sample is left out in the *CV-Samples* procedure, during testing these cells are rejected because they are missing in the training data. These results support the validity of using the rejection option to label unknown cells, which are likely to be rare sample-specific populations.

5.3.8 LINEAR CLASSIFICATION IS SUFFICIENT FOR ACCURATE CLASSIFICATION OF CYTOF DATA

We have shown that a simple linear classifier such as LDA has a better performance compared to complex non-linear classifiers such as ACDC and DeepCyTOF. To further illustrate that non-linear classification does not perform better than linear classification, we compared the performance of LDA to a k-NN classifier on the HMIS-2 dataset. We found that LDA has a comparable performance to a k-NN classifier with k = 50 (Table 5.2), suggesting that adding non-linearity to the classification process does not improve performance.

Further, we checked the effect of having similar populations on the classification performance. For each cell population in the HMIS-2 dataset, we compared the F1-score with the correlation to the most similar population (Supplementary Figure 5.10). For both, LDA and k-NN classifiers, we observe a week negative relation, showing that the classifier performance is affected by the presence of similar cell populations in the dataset.

To reduce the computation time for the k-NN classifier, we employed an editing scheme to reduce the size of the training data (see Methods). Using the proposed editing scheme, we reduced the training data size to an average of 300,000 per training fold (~12 % of the original training set), resulting in a significant speedup of the training and testing times. However, the k-NN classifier still takes on average 180x the time needed by LDA to make predictions for one sample.

Next, we investigated whether feature selection (using less markers during classification) would affect the performance of the classifiers. The k-NN classifier selected only 20 (out of the 28) markers and retained a comparable performance to that obtained using all 28 markers. On the other hand, feature selection did not reduce the number of markers selected by LDA, indicating that LDA requires all the measured markers in order to achieve maximum performance.

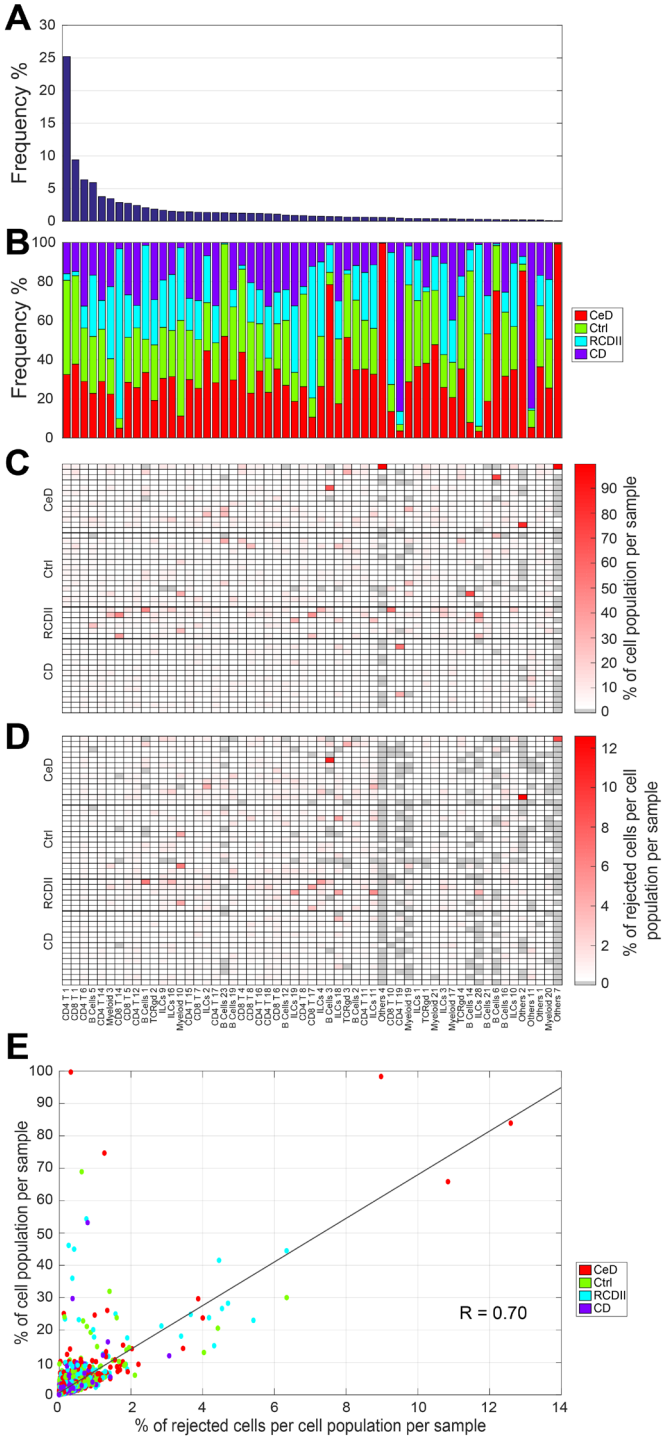


Figure 5.4 Rejection option effect on variable sized cell populations. (**A**) Cell population frequency across the HMIS-2 dataset, in a descend order. (**B**) Cell population composition in terms of the different

sample types (CeD, Ctrl, RCDII, and CD). (**C**) Cell population frequencies across samples, normalized by the cell population size across all samples, every column summation is 100%. (**D**) Percentage of rejected cells per cell population per sample, normalized by the cell population size across all samples, using a posterior probability threshold of 0.7. Cell populations follow the same order for (**A**–**D**). (**E**) Scatter plot between values in (**C**) and (**D**) showing a strong correlation of 0.70 between the rejection ratio and the cell population size, per sample. Each point represents a cell population in a particular sample, and points are colored according to the disease status of the sample annotation.

5.4 DISCUSSION

In this work, we showed that a linear classifier can be used to automatically assign labels to single cells in mass cytometry data. Using four different CyTOF datasets, we compared the performance of a linear discriminant analysis classifier (LDA) to two recent approaches methods: ACDC[22] and DeepCyTOF[23]. Interestingly, LDA has better performance compared to ACDC and DeepCyTOF in all four datasets. Compared to ACDC, LDA does not require any additional biological knowledge or assumptions regarding the distribution patterns of markers. Additionally, ACDC requires a cell-type marker table which has several limitations: (i) designing the table can be very challenging in the presence of many cell populations, (ii) it is not possible to specify the marker patterns for some cell populations (e.g. ACDC ignored 4 subtypes in the BMMC dataset because the table could not be constructed), and (iii) the table requires imposing assumptions on the marker distribution (currently binary) which can be challenging to model. Furthermore, results on the BMMC dataset show that LDA can detect rare cell populations having frequencies < 0.5% of the total number of cells, like MPP, HSC, MEP and GMP, which were the main cause of the lower performance of ACDC[22]. Compared to DeepCyTOF, in addition to have better performance, LDA is a much simpler classifier which means it has substantial advantages with respect to the interpretability of the classifier prediction, reproducibility, and scalability to larger datasets with deep subtyping annotation.

We further evaluated LDA on a large CyTOF dataset with deep annotation of cell populations. We showed that LDA can accurately identify cell populations in a challenging dataset of 3.5 million cells comprised of 57 cell populations. Further, we showed that the errors made by LDA in assigning cell population labels to each cell has negligible influence on the estimates of cell population frequencies across different individuals. DeepCyTOF failed to scale, in terms of performance, to this large dataset with deep level of annotation. Its low performance is mainly due to the selection of one sample for training. Moreover, this approach is particularly not suitable when analysing multiple samples from different cohorts (e.g. disease and controls). For instance, in the HMIS-2 dataset, DeepCyTOF selected sample (number 27) as the training sample, which is a control sample containing only 55 of the 57 cell populations.

We also compared LDA to a simpler classifier such as the NMC, to test to which extend the classification task could be further simplified. We observed comparable performance in datasets containing large and major cell populations only, such as Multi-Center and HMIS-1, where the classification task is relatively easy. However, LDA produces better results for other datasets, having more detailed population subtyping, in which the classification task becomes more challenging, and NMC performance drops, especially for small populations as observed in the BMMC dataset.

To show that a linear classifier is sufficient to classify cells in mass cytometry data, we compared LDA to a non-linear classifier (k-NN). Indeed, the k-NN classifier does not outperform LDA on the HMIS dataset, indicating that there is no added value in using non-linear relationships between the markers. However, when we ran both classifiers with feature selection, LDA required the full set of markers to achieve the best performance. On the other

hand, the k-NN classifier was able to achieve the same performance as LDA but using less markers (20 instead of 28). This result suggests that a non-linear classifier might be beneficial to reduce the number of required markers and free valuable slots on the CyTOF panel for additional markers. Alternatively, using the reduced marker set lowers costs when analysing new samples, using a smaller CyTOF panel or even flow cytometry while retaining the ability to identify all cell populations of interest.

Further, the comparable performance of LDA and k-NN indicate that in the full marker space, the cell population classes in the CyTOF datasets that we explored are well separable. Consequently, different clustering algorithms will perform similarly well on these datasets. We would like to note that more complex data might need more complex classifiers or clustering algorithms, for example when cell populations are less separable like continuous or smeary populations. We have shown that for the current datasets this is not necessary. In general, it will be difficult to predict beforehand which complexity is necessary, so that in practice multiple classifiers need to be evaluated.

Our results also show that the performance of LDA is not largely affected by either technical or biological variability. Technical variability is part of the Multi-Center dataset which contains batch effects. The performances on the different batch samples remain relatively high (weighted F1-score >0.95, Supplementary Figure 5.3), although, applying batch correction methods might still improve the overall LDA prediction performance[29–31]. Biological variability is presented in the HMIS dataset, which includes samples from patients with different diseases, collected over time. The high performance on the deeply annotated HMIS-2 dataset, shows LDA's robustness against these biological variations.

For the HMIS dataset, we relied on an initial clustering step to assign ground-truth labels. To avoid any possible leakage of information from the test set of cells by including them into the clustering, we designed a conservative learning scheme. In the conservative scheme, we don't use the labels obtained by clustering the entire dataset (i.e. ground-truth) for training, but rather re-cluster the training data inside each fold. In addition, this scheme better resembles a realistic scenario in which the new unseen data is never included in the initial assignment of class labels for training. The performance of LDA in this conservative experiment is lower than the initial performance obtained by classical cross validation. However, the lower performance does not stem from the lack of generalization, as the results show high performance on the overview-level, but rather from the difficulty in matching cluster labels between the ground truth and the training set.

Clustering approaches in general have an advantage over classification methods in that they can be employed to discover new cell populations. However, an additional advantage of using a probabilistic classifier such as LDA is that we can directly gain information regarding the accuracy of each decision made by inspecting the posterior probability. We showed that we can allow for a rejection option when the posterior probability of the classification of a particular cell is low. This rejection option can be used to identify "unknown" cells which might require additional investigation to determine their biological relevance. Additionally, we showed that these 'unknown' cells are likely to be rare and sample-specific. There is however a trade-off between how confident we are on the correctness of the predictions and the size of the 'unknown' class. A stringent threshold (i.e. high posterior probability) means that many cells will be classified as 'unknown' which will further require manual investigation.

Taken together, we demonstrated the feasibility of using a simple linear classifier to automatically label cells in mass cytometry data which is a promising step forward to use mass cytometry data in cohort studies.

5.5 AVAILABILITY

Data is available from Flow Repository (FR-FCM-ZYTT) and implementation is available on GitHub (https://github.com/tabdelaal/CyTOF-Linear-Classifier).

BIBLIOGRAPHY

1. Bandura, D. R. *et al.* Mass Cytometry : Technique for Real Time Single Cell Multitarget Immunoassay Based on Inductively Coupled Plasma Time-of-Flight Mass Spectrometry. *Anal. Chem.* **81,** 6813–6822 (2009).
2. Spitzer, M. H. & Nolan, G. P. Mass Cytometry: Single Cells, Many Features. *Cell* **165,** 780–791 (2016).
3. Amir, E. D. *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31,** 545–552 (2014).
4. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162,** 184–197 (2015).
5. Chevrier, S. *et al.* An Immune Atlas of Clear Cell Renal Cell Carcinoma. *Cell* **169,** 736–749 (2017).
6. van Unen, V. *et al.* Mass Cytometry of the Human Mucosal Immune System Identifies Tissue- and Disease-Associated Immune Subsets. *Immunity* **44,** 1227–1239 (2016).
7. Newell, E. W., Sigal, N., Bendall, S. C., Nolan, G. P. & Davis, M. M. Cytometry by Time-of-Flight Shows Combinatorial Cytokine Expression and Virus-Specific Cell Niches within a Continuum of CD8 + T Cell Phenotypes. *Immunity* **36,** 142–152 (2012).
8. Newell, E. W. *et al.* Combinatorial tetramer staining and mass cytometry analysis facilitate T-cell epitope mapping and characterization. *Nat. Biotechnol.* **31,** 623–629 (2013).
9. Newell, E. W. & Cheng, Y. Mass cytometry: Blessed with the curse of dimensionality. *Nature Immunology* **17,** 890–895 (2016).
10. Qiu, P. *et al.* Extracting a Cellular Hierarchy from High-dimensional Cytometry Data with SPADE. *Nat. Biotechnol.* **29,** 886–891 (2012).
11. Van Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytom. Part A* **87,** 636–645 (2015).
12. Samusik, N., Good, Z., Spitzer, M. H., Davis, K. L. & Nolan, G. P. Automated mapping of phenotype space with single-cell data. *Nat. Methods* **13,** 493–496 (2016).
13. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24,** 417–441 (1933).
14. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn.* **9,** 2579–2605 (2008).
15. Pezzotti, N. *et al.* Approximated and User Steerable tSNE for Progressive Visual Analytics. *IEEE Trans. Vis. Comput. Graph.* **23,** 1739–1752 (2017).
16. Pezzotti, N., Höllt, T., Lelieveldt, B., Eisemann, E. & Vilanova, A. Hierarchical Stochastic Neighbor Embedding. in *Computer Graphics Forum (Proceedings of EuroVis 2016)* **35,** (2016).
17. Van Unen, V. *et al.* Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. *Nat. Commun.* **8,** 1–10 (2017).
18. Weber, L. M. & Robinson, M. D. Comparison of Clustering Methods for High-Dimensional Single-Cell Flow and Mass Cytometry Data. *Cytom. A* **89,** 1084-1–96 (2016).
19. Maecker, H. T., McCoy, J. P. & Nussenblatt, R. Standardizing immunophenotyping for the Human Immunology Project. *Nature Reviews Immunology* **12,** 191–200 (2012).
20. Hsiao, C. *et al.* Mapping cell populations in flow cytometry data for cross-sample comparison using the Friedman-Rafsky test statistic as a distance measure. *Cytom. Part A* **89,** 71–88 (2016).
21. Lux, M. *et al.* flowLearn : Fast and precise identification and quality checking of cell

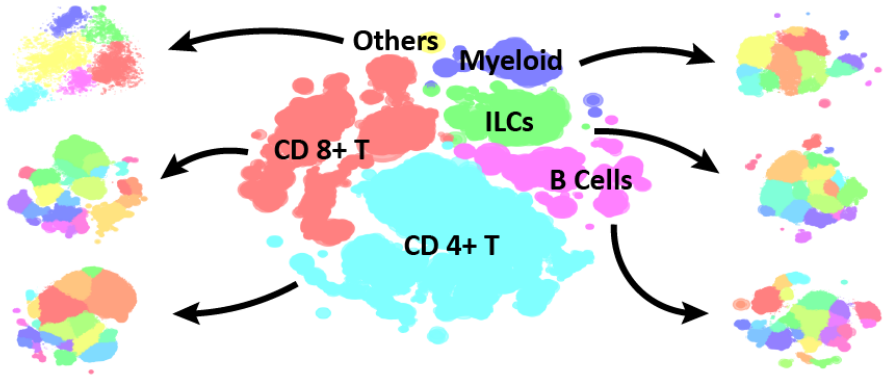
populations in flow cytometry. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty082/4860364
22. Lee, H., Kosoy, R., Becker, C. E., Dudley, J. T. & Kidd, B. A. Automated cell type discovery and classification through knowledge transfer. *Bioinformatics* **33,** 1689–1695 (2017).
23. Li, H. *et al.* Gating mass cytometry data by deep learning. *Bioinformatics* **33,** 3423–3430 (2017).
24. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34,** 1145–1160 (2016).
25. Bendall, S. C. *et al.* Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science (80-.).* **332,** 687–696 (2011).
26. Höllt, T. *et al.* Cytosplore : Interactive Immune Cell Phenotyping for Large Single-Cell Datasets. in *Computer Graphics Forum (Proceedings of EuroVis 2016)* **35,** (2016).
27. Comaniciu, D. & Meer, P. Mean Shift : A Robust Approach Toward Feature Space Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **24,** 603–619 (2002).
28. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36,** 89–94 (2018).
29. Shaham, U. *et al.* Removal of batch effects using distribution-matching residual networks. *Bioinformatics* **33,** 2539–2546 (2017).
30. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36,** 411–420 (2018).
31. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36,** 421–427 (2018).

SUPPLEMENTARY MATERIALS

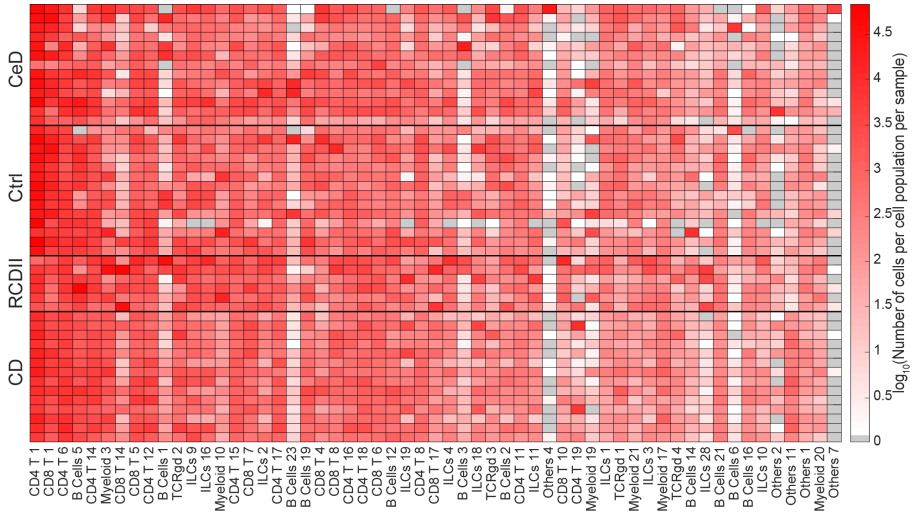
Supplementary Table 5.1 Summary of the datasets used in this study

Dataset	Samples	Cells	Markers	Cell populations
AML	n.a.	104,184	32	14
BMMC	n.a.	81,747	13	24
PANORAMA	10	514,386	39	24
Multi-Center	16	929,685	8	5
HMIS-1	47	3,553,596	28	6
HMIS-2	47	3,553,596	28	57

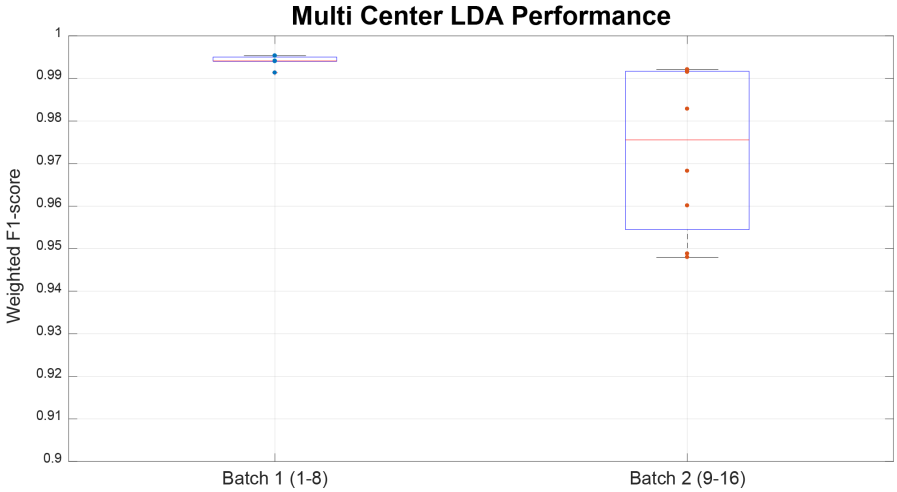
n.a. = not available



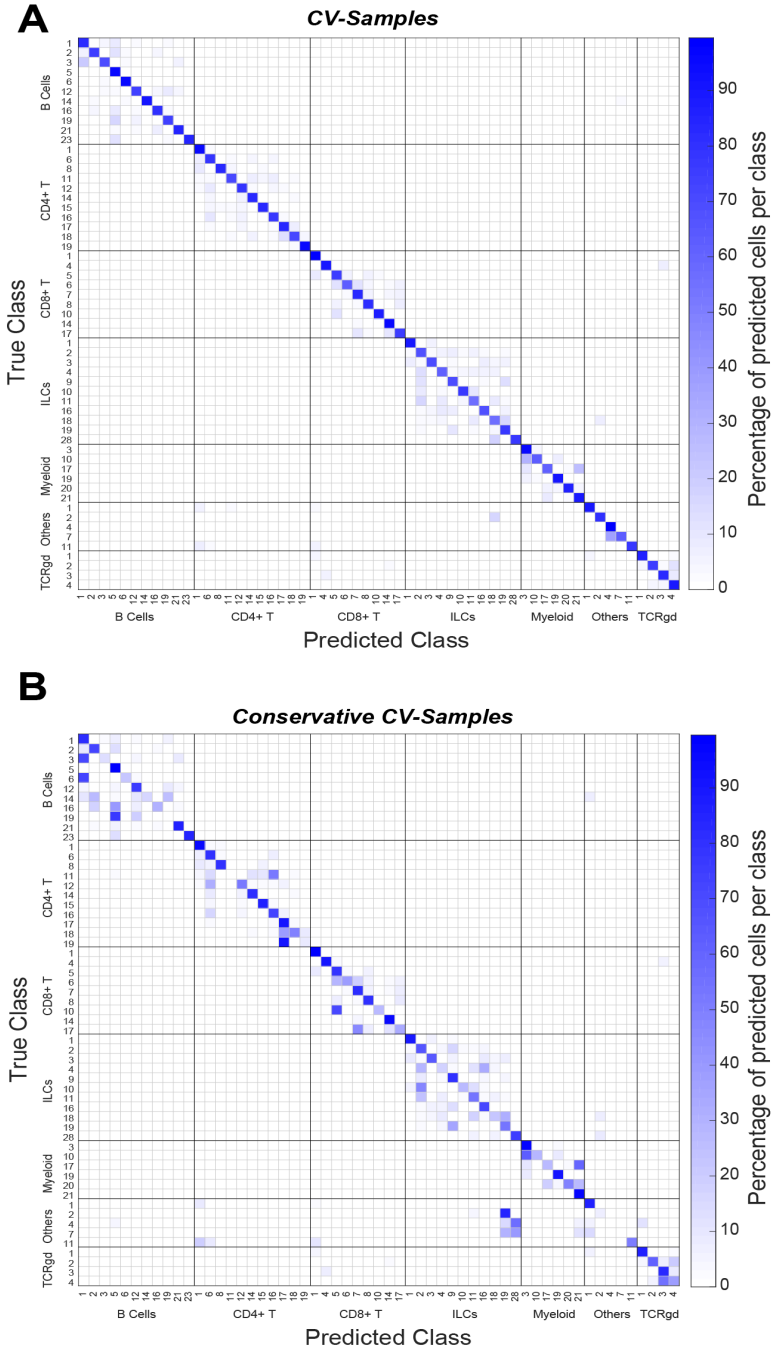
Supplementary Figure 5.1 Annotation of cells in the HMIS dataset. The middle image shows the embedding of the overview (top) HSNE layer, clustered into six major cell populations. Next, a separate tSNE map is obtained per cell population by exploring one layer down.



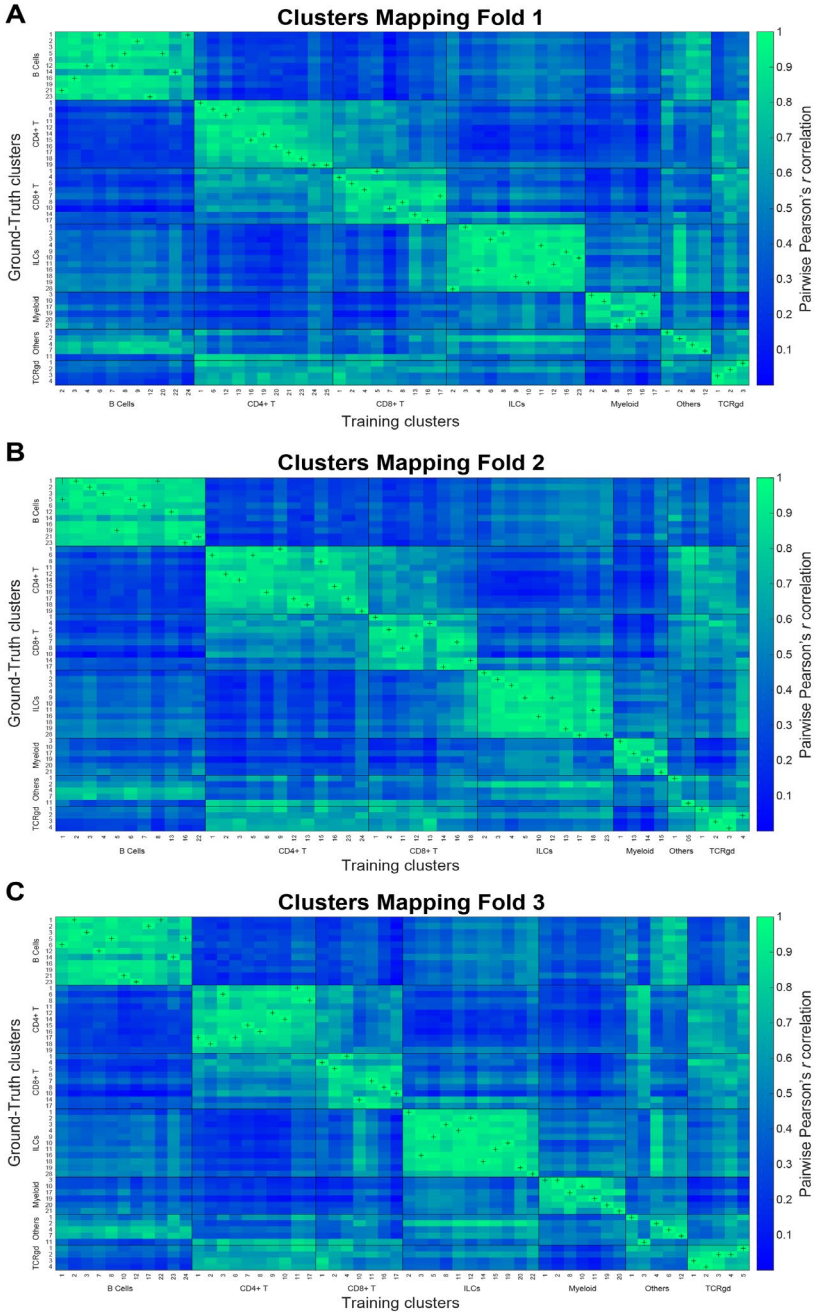
Supplementary Figure 5.2 Cell counts per cell population per sample for the HMIS-2 dataset. Cell populations are ordered in descending order of the frequencies across all samples, and grouped according to the sample types (CeD = Celiac Disease, Ctrl = Control, RCDII = Refractory Celiac Disease Type II, CD = Crohn's Disease). All counts were $\log_{10}$ transformed.



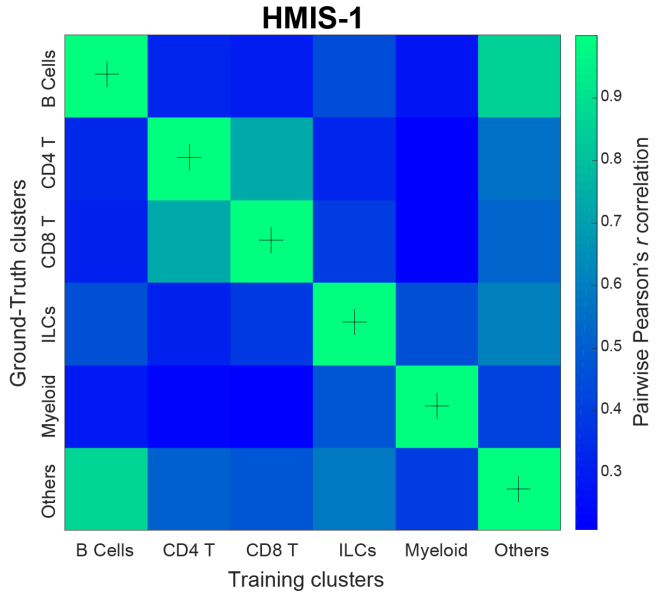
Supplementary Figure 5.3 LDA performance for the Multi-Center dataset, using only sample 2 as training and calculating the weighted F1-score for the 15 remaining samples. Batch 1 is sample 1-8, and batch 2 is sample 9-16.



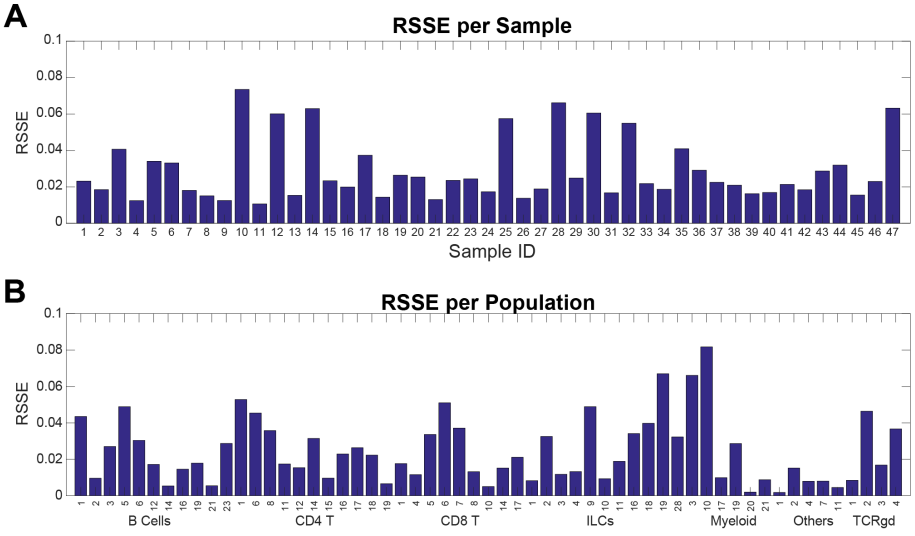
Supplementary Figure 5.4 LDA performance on the HMIS-2 dataset. (A) Classification confusion matrix when using *CV-Samples* setup, showing high percentages along the matrix diagonal, as well as that most of the misclassification (off-diagonal values) falls within the major cell populations. **(B)** Classification confusion matrix when using *Conservative CV-Samples* setup, showing lower percentages along the matrix diagonal compared to the *CV-Samples* setup. Each cell (square) in the confusion matrix represents the percentage of overlapping cells between true and predicted class.



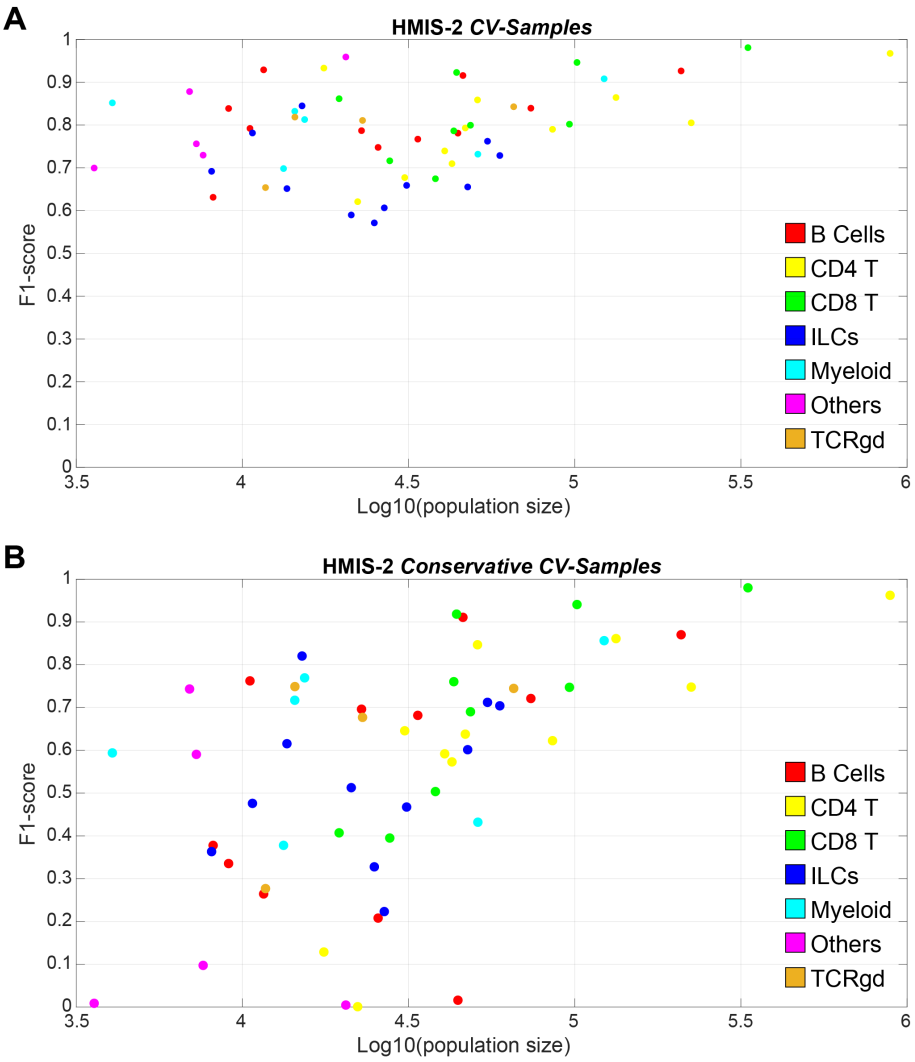
Supplementary Figure 5.5 Mapping of training clusters to ground-truth clusters during the *Conservative CV-Samples* setup of HMIS-2 dataset. **(A-C)** correlation maps for all three folds, highlighting the maximum correlation with a '+' sign.



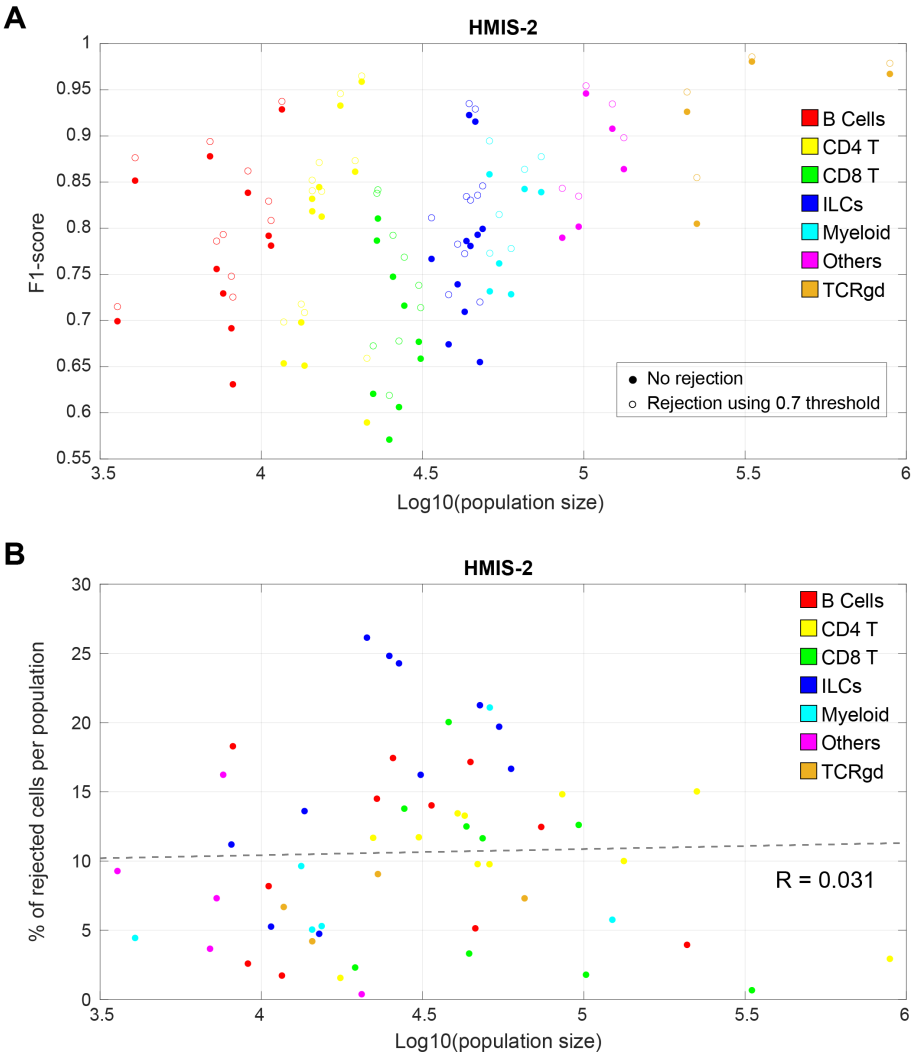
Supplementary Figure 5.6 Mapping of training clusters to ground-truth clusters during the *Conservative CV-Samples* setup of HMIS-1 dataset, highlighting the maximum correlation with a '+' sign.



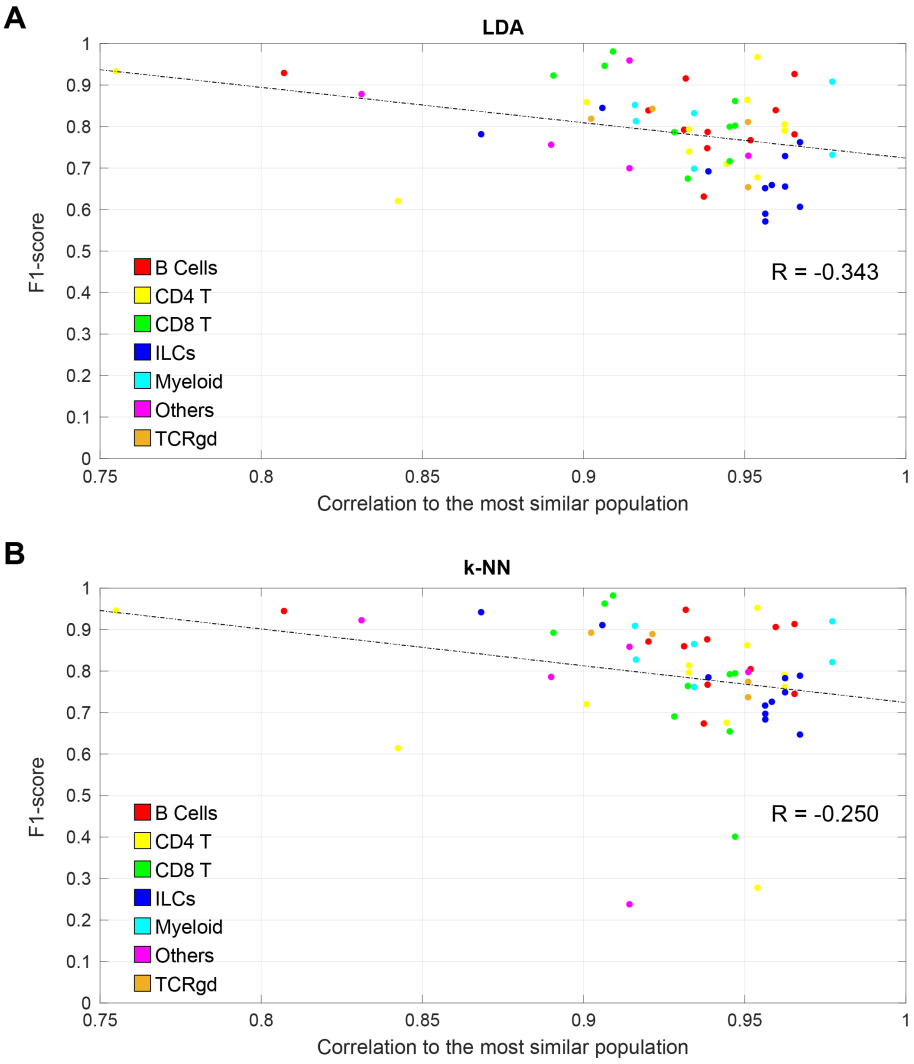
Supplementary Figure 5.7 Bar plot of the Root of Sum Squared Error (RSSE) **(A)** per sample, and **(B)** per cell population.



Supplementary Figure 5.8 Relationship between performance and population size. Scatter plot of the F1-score vs. the population size for the HMIS-2 dataset evaluated using **(A)** *CV-Samples*, and **(B)** *Conservative CV-Samples*. Each dot represents one cell population and coloured according to the major cell population annotation.



Supplementary Figure 5.9 (A) Cell populations F1-score with and without rejection, using a rejection threshold of 0.7, **(B)** Scatter plot between the population size and the percentage of rejected cells per population, showing no correlation ≈ 0.



Supplementary Figure 5.10 Scatter plots showing the F1-score per population vs the correlation of the most similar population in the HMIS-2 dataset, for **(A)** LDA classifier, and **(B)** k-NN classifier. In both classifier, we observed a week negative correlation.

CHAPTER 6

HIGH-DIMENSIONAL CYTOMETRIC ANALYSIS OF COLORECTAL CANCER REVEALS NOVEL MEDIATORS OF ANTITUMOUR IMMUNITY

Natasja L. de Vries

Vincent van Unen[*]

Marieke E. Ijsselsteijn[*]

Tamim Abdelaal[*]

Ruud van der Breggen

Arantza Farina Sarasqueta

Ahmed Mahfouz

Koen C.M.J. Peeters

Thomas Höllt

Boudewijn P.F. Lelieveldt

Frits Koning[#]

Noel F.C.C. de Miranda[#]

This chapter is published in: *Gut* (2019) 69(4): 691-703, doi: 10.1136/gutjnl-2019-318672.
Supplementary material is available online at: https://gut.bmj.com/content/69/4/691.full
[*,#]Equal contribution

A comprehensive understanding of anti-cancer immune responses is paramount for the optimal application and development of cancer immunotherapies. We unraveled local and systemic immune profiles in colorectal cancer (CRC) patients by high-dimensional analysis to provide an unbiased characterization of the immune contexture of CRC. Thirty-six immune cell markers were simultaneously assessed at the single-cell level by mass cytometry in 35 CRC tissues, 26 tumour-associated lymph nodes, 17 colorectal healthy mucosa, and 19 peripheral blood samples from 31 CRC patients. Additionally, functional, transcriptional, and spatial analyses of tumour-infiltrating lymphocytes were performed by flow cytometry, single-cell RNA-sequencing, and multispectral immunofluorescence.

We discovered that a previously unappreciated innate lymphocyte population (Lin-CD7+CD127-CD56+CD45RO+) was enriched in CRC tissues and displayed cytotoxic activity. This subset demonstrated a tissue-resident (CD103+CD69+) phenotype, and was most abundant in immunogenic mismatch repair (MMR)-deficient CRCs. Their presence in tumours was correlated with the infiltration of tumour-resident cytotoxic, helper, and γδ T cells with highly similar activated (HLA-DR+CD38+PD-1+) phenotypes. Remarkably, activated γδ T cells were almost exclusively found in MMR-deficient cancers. Non-activated counterparts of tumour-resident cytotoxic and γδ T cells were present in CRC and healthy mucosa tissues, but not in lymph nodes, with the exception of tumour-positive lymph nodes.

In conclusion, this work provides a blueprint for the understanding of the heterogeneous and intricate immune landscape of CRC, including the identification of previously unappreciated immune cell subsets. The concomitant presence of tumour-resident innate and adaptive immune cell populations suggests a multi-targeted exploitation of their anti-tumour properties in a therapeutic setting.

6.1 INTRODUCTION

T cell checkpoint blockade immunotherapies have revolutionized cancer treatment following the clinical success achieved with therapeutic antibodies targeting CTLA-4 and the PD-1/PD-L1 axis in cancer patients. These strategies reinvigorate anti-tumour T cell responses, and are particularly effective in cancers with high mutation burden like melanomas, non-small cell lung cancers, and DNA mismatch repair (MMR)-deficient cancers[1–5]. MMR deficiency occurs in approximately 15-20% of colorectal cancers (CRCs) and leads to the widespread accumulation of somatic mutations in tumours, including insertions and deletions at DNA microsatellite sequences[6,7]. Such a theoretically immunogenic profile is corroborated by the presence of numerous intraepithelial lymphocytes in these cancers, in contrast to MMR-proficient cancers[8,9]. Nevertheless, not all MMR-deficient CRCs respond to immune checkpoint blockade, while MMR-proficient CRCs are insensitive to this therapy.

To understand the mechanisms that determine responses to current immunotherapies and for the design of alternative approaches, it is crucial to characterize the cancer microenvironment with multidimensional approaches that allow the simultaneous identification and characterization of immune cell populations across multiple lineages[10,11]. Mass cytometry allows a detailed single-cell characterization of adaptive and innate immune landscapes, thereby providing a unique platform to discriminate immune cell subsets that can be exploited in an immunotherapeutic setting.

We performed an in-depth characterization of immune landscapes across CRC tissues, tumour-associated lymph nodes, colorectal healthy mucosa, and peripheral blood samples from 31 CRC patients by high-dimensional single-cell mass cytometry. We revealed tumour

tissue-specific immune signatures across the adaptive and innate compartments, and discovered a previously unappreciated innate immune cell population implicated in anti-tumour immunity that strongly differentiated immunogenic (MMR-deficient) from non-immunogenic (MMR-proficient) CRCs.

6.2 MATERIALS AND METHODS

6.2.1 HUMAN SAMPLES

Primary CRC tissues (N=35, of which 22 MMR-proficient and 13 MMR-deficient) with matched tumour-associated lymph nodes (N=26), colorectal healthy mucosa (N=17), and pre-surgical peripheral blood samples (N=19) from 31 CRC patients were processed for this study (Supplementary Table 6.1). All patients were treatment-naïve except five rectal cancer patients which received neo-adjuvant therapy (Supplementary Table 6.1). One patient was diagnosed with multiple primary colorectal tumours (N=5) at different locations, all of which were included in the study (Supplementary Table 6.1). No patient with a previous history of inflammatory bowel disease was studied. To account for tumour heterogeneity, macroscopic sectioning from the lumen to the most invasive area of the tumour was performed for further processing. This study was approved by the Medical Ethical Committee of the Leiden University Medical Center (protocol P15.282), and patients provided written informed consent. All specimens were anonymized and handled according to the ethical guidelines described in the Code for Proper Secondary Use of Human Tissue in the Netherlands of the Dutch Federation of Medical Scientific Societies.

6.2.2 TISSUE PROCESSING AND MASS CYTOMETRY ANTIBODY STAINING

Details on tissue processing and mass cytometry antibody staining are available in online Supplementary Methods and Supplementary Table 6.2.

6.2.3 MASS CYTOMETRY DATA ANALYSIS

Mass cytometry experiments were performed with a discovery and validation cohort of CRC patients. The discovery cohort consisted of 19 CRC tissues, 17 tumour-associated lymph nodes, 4 colorectal healthy mucosa, and 9 peripheral blood samples. Single, live $CD45^+$ cells were gated in Cytobank[12] (Supplementary Figure 6.1). $CD45^+$ cells were sample-tagged, hyperbolic ArcSinh transformed with a cofactor of 5, and subjected to dimensionality reduction analysis in Cytosplore[13] Of the 39 antibodies included in the panel, 36 showed clear discrimination between positive and negative cells (Supplementary Figure 6.1). Major immune lineages (Figure 6.1A-B) were identified at the overview level of a 5-level Hierarchical Stochastic Neighbour Embedding (HSNE) analysis[14,15] on $CD45^+$ data from all samples ($8.9*10^6$ cells) with default perplexity and iterations (30 and 1,000, respectively). Naive and memory $CD4^+$ and $CD8^+/\gamma\delta$ T cell, B cell, Lin^-CD7^+ innate lymphoid cell (ILC), and myeloid cell lineages were analyzed in a data-driven manner up to a maximum number of $0.5*10^6$ landmarks[15]. Clustering of the data was performed by Gaussian Mean Shift (GMS) clustering in Cytosplore, and an algorithm was run that merged clusters showing high similarity in ArcSinh5-transformed median expression of all markers (<1). Hierarchical clustering on cell frequencies was performed in Matlab using Spearman's rank correlation.

The validation cohort consisted of 16 CRC tissues, 9 tumour-associated lymph nodes, 13 colorectal healthy mucosa, and 10 peripheral blood samples. Single, live $CD45^+$ cells were hyperbolic ArcSinh transformed with a cofactor of 5, and classified into the pre-identified immune cell clusters of the discovery cohort based on similarity in marker expression. To

obtain consistent cell clusters across both cohorts, a Linear Discriminant Analysis classifier was trained using the cell clusters of the discovery cohort and was used to automatically predict the cluster label for each cell in the validation cohort[16]. To account for technical variation, a peripheral blood mononuclear cell (PBMC) reference sample was included in every mass cytometry experiment. ComBat was applied to align the PBMC reference samples and corresponding patient samples to correct for batch effects[17].

6.2.4 SINGLE-CELL RNA-SEQUENCING

$CD45^+$ cells from 7 tumours (4 MMR-deficient and 3 MMR-proficient) were MACS-sorted with anti-CD45-PE antibodies (clone 2D1, Thermo Fisher Scientific) and anti-PE microbeads (Miltenyi Biotec). Single-cell RNA-sequencing libraries were prepared using the Chromium Single Cell 3' Reagent Kit, Version 2 Chemistry (10x Genomics) according to the manufacturer's protocol. Libraries were sequenced on a NovaSeq6000 using paired-end 2x150bp sequencing (Illumina). Downstream analysis was performed using the Seurat R package according to the author's instructions[18]. Briefly, cells with fewer than 200 expressed genes, and genes that were expressed in less than 3 cells were excluded. Furthermore, cells with outlying percentages of differentially expressed mitochondrial genes (>0.20) and cells with outlying numbers of expressed genes (>5000) were excluded. This resulted in a final dataset of 1,079 cells expressing a total of 1,972 variable genes. Cells were pre-processed using principal component analysis, clustered using graph-based community detection[19], and visualized by t-distributed Stochastic Neighbour Embedding (t-SNE)[20]. Differentially expressed genes were identified for each cell cluster and visualized in violin plots. In addition, $CD45^+$ cells from one MMR-deficient tumour with high numbers of $Lin^-CD7^+CD127^-CD56^+CD45RO^+$ ILCs were sorted on a FACS Aria II sorter (BD Biosciences) (Supplementary Table 6.3). A similar single-cell RNA-sequencing analysis pipeline was performed while sequencing was performed on a HiSeq4000 (Illumina). Cut-offs for outlying percentages of differentially expressed mitochondrial genes (>0.05) and cells with outlying numbers of expressed genes (>5500) were used. Here, a final dataset of 795 cells expressing a total of 1,814 variable genes was obtained.

6.2.5 FLOW CYTOMETRY

Single-cell suspensions of CRC tissues (N=8, of which 5 MMR-deficient and 3 MMR-proficient) were stimulated in IMDM/L-glutamine medium (Lonza) complemented with 10% human serum with 20 ng/mL PMA (Sigma-Aldrich) and 1 μg/mL ionomycin (Sigma-Alrdich) for 6 hr at 37°C. Ten μg/mL brefeldin A (Sigma-Aldrich) was added for the last 4 hours. A flow cytometry antibody panel was designed to detect granzyme B/perforin, IFN-γ, and TNF-α production by ILC, T cell, and γδ T cell populations (Supplementary Table 6.3). In addition, FOXP3 expression by $ICOS^+$ regulatory T cells was assessed in single-cell suspensions of CRC tissues (N=4, of which 1 MMR-deficient and 3 MMR-proficient). Details on flow cytometry antibody staining are available in online supplementary methods.

6.2.6 IMMUNOHISTOCHEMICAL STAINING

Details on immunohistochemical detection of MMR proteins and human leukocyte antigen (HLA) class I expression of CRC tissues are available in online Supplementary Methods.

6.2.7 MULTISPECTRAL IMMUNOFLUORESCENCE

A six-marker immunofluorescence panel was applied to 5-μm frozen tissue sections of 4 MMR-deficient and 4 MMR-proficient colorectal tumours, as described previously[21]. Details on immunofluorescence antibody staining are available in online Supplementary Methods and

Supplementary Table 6.4. For each tumour, five different tissue sections were imaged at 20x magnification with the Vectra 3.0 Automated Quantitative Pathology Imaging System (Perkin Elmer). InForm Cell Analysis software (Perkin Elmer) was used for image analysis and spectral separation of dyes, by using spectral libraries defined with single-marker immunofluorescence detection. Tissue segmentation was trained manually with DAPI to segment images into tissue and 'no tissue' areas. All images were visually inspected for the number of $CD3^-TCR\alpha\beta^-CD127^-CD7^+CD45RO^+$ ILCs and cell counts were normalized by tissue area (number of cells per mm^2).

6.2.8 STATISTICAL ANALYSIS

Data were presented as median ± interquartile range. Group comparisons were performed with Mann-Whitney U test, Kruskal-Wallis test with Dunn's test for multiple comparisons, or Friedman test with Dunn's test for multiple comparisons (GraphPad Prism version 7), as indicated. In the correlation analysis, *P*-values were adjusted for multiple testing using the Benjamini-Hochberg procedure. *P*-values < 0.05 were considered statistically significant.

6.3 RESULTS

6.3.1 TUMOR-RESIDENT IMMUNE CELL POPULATIONS DERIVE FROM MULTIPLE LINEAGES

Mass cytometric analysis of 36 immune cell markers was performed on single-cell suspensions isolated from cancer and healthy tissues of CRC patients. To decipher their immune composition, we performed HSNE analysis in Cytosplore on all acquired $CD45^+$ cells of the discovery cohort ($8.9*10^6$ cells in total) (Figure 6.1A). Based on the density features of the HSNE-embedded landmarks, 7 major immune cell clusters were identified by unsupervised GMS clustering, which corresponded to naive and memory (based on CD45RO and CCR7 expression) $CD4^+$ and $CD8^+/\gamma\delta$ T cells, B cells, Lin^-CD7^+ ILCs, and myeloid cells (Figure 6.1A-B). Memory $CD4^+$ and $CD8^+$ T cells, as well as myeloid cells, were dominant immune lineages in the tumour microenvironment, while B cells, Lin^-CD7^+ ILCs, and naive $CD4^+$ and $CD8^+$ T cells were present at a lower extent (Supplementary Figure 6.2). The HSNE analysis also unveiled the presence of several tumour tissue-specific, phenotypically distinct landmarks within the memory $CD4^+$ T cell, $CD8^+/\gamma\delta$ T cell, Lin^-CD7^+ ILC, and myeloid cell compartments (Figure 6.1A).

All 7 major immune lineages were analyzed in detail by hierarchical exploration of the data in HSNE. As an example, the embedding of the memory $CD8^+/\gamma\delta$ T cell compartment is shown in Figure 6.1C. Altogether, analysis of these 7 major immune lineages yielded 220 distinct immune cell clusters, of which 2 consisted of less than 100 cells and were excluded from further analysis. All acquired $CD45^+$ cells of the validation cohort ($6.6*10^6$ cells in total) were subsequently classified into these pre-identified immune cell clusters based on their phenotype (see Methods).

The mass cytometric analysis was accompanied by single-cell RNA-sequencing of $CD45^+$ cells from 7 CRC tissues. Seven immune cell clusters could be detected based on transcriptomic profiles (Figure 6.1D), corresponding to B cells, $CD8^+$ and $CD4^+$ T cells, ILCs, myeloid cells, proliferating cells, and plasma B cells (Figure 6.1D-E).

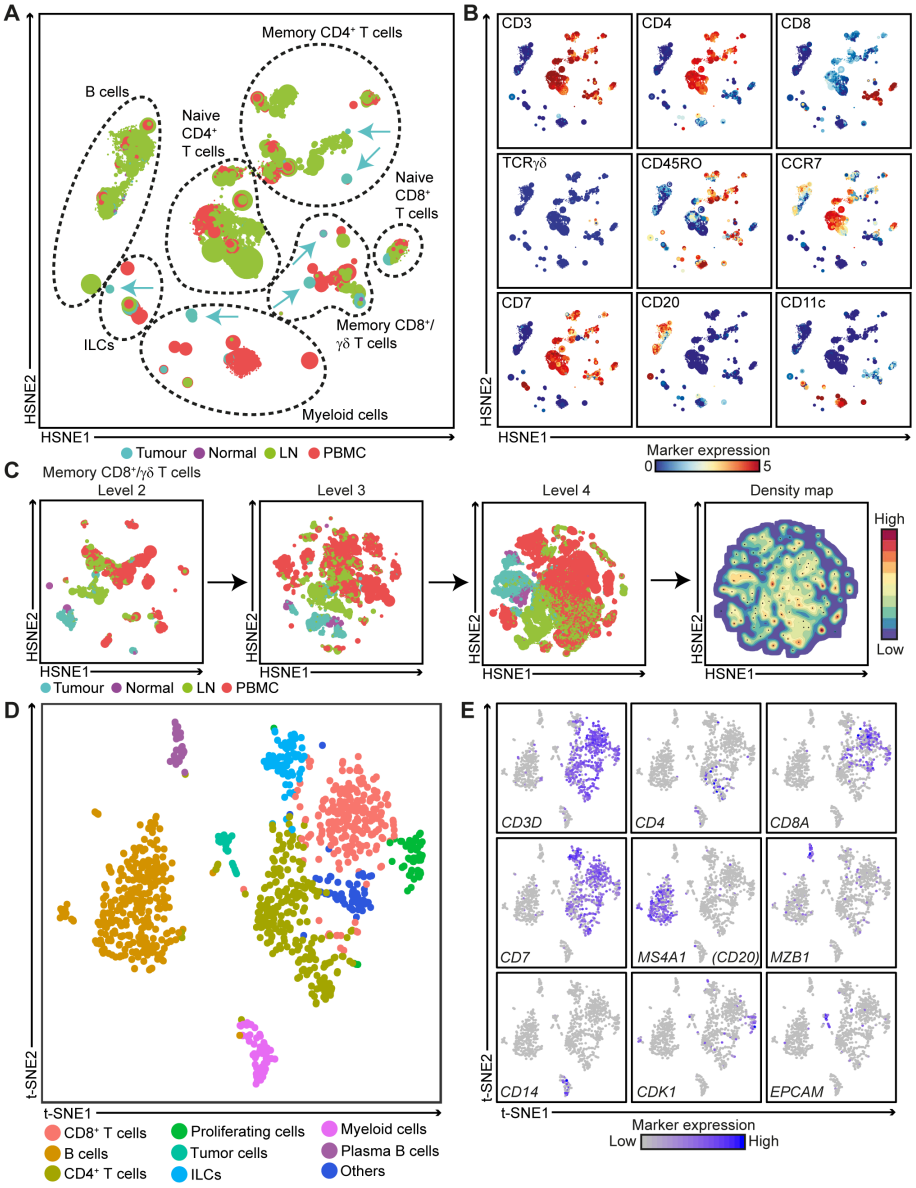


Figure 6.1 Tumour-resident immune cell populations derive from multiple lineages. (A and B) HSNE embedding showing $7.5*10^4$ landmarks representing immune cells ($8.9*10^6$ cells) isolated from CRC tissues (N=19), tumour-associated lymph nodes (N=17), colorectal healthy mucosa (N=4), and peripheral blood (N=9) samples from the discovery cohort. Colors represent the different tissue types **(A)** and the relative expression of indicated immune lineage markers **(B)**. Arrows indicate the HSNE location of phenotypically distinct tumour-resident immune cell populations. **(C)** Example of an HSNE analysis of $7.4*10^2$ landmarks representing $1.1*10^6$ cells from the memory CD8⁺/γδ T cell compartment as identified in (A). All landmarks are selected and embedded at the next, more detailed levels showing a finer granularity of structures with $5.0x10^3$ landmarks at level 2, to $3.0*10^4$ landmarks at level 3, and $1.6*10^5$ landmarks at level 4. Phenotypically distinct immune cell clusters were identified by unsupervised GMS clustering based on the density features. Black dots indicate the centroids of the identified clusters. **(D and E)** t-SNE embedding showing 1,079 cells from CRC tissues (N=7) analyzed by single-cell RNA-

sequencing. Colors represent the different clusters **(D)** and the log-transformed expression levels of indicated immune lineage markers **(E)**. Each dot represents a single cell. GMS; Gaussian mean shift, HSNE; hierarchical stochastic neighbour embedding, LN; lymph node, PBMC; peripheral blood mononuclear cell, t-SNE; t-distributed stochastic neighbour embedding.

6.3.2 ACTIVATED CD8$^+$ AND γδ T CELLS ARE TUMOUR TISSUE-SPECIFIC AND ENRICHED IN MISMATCH REPAIR-DEFICIENT COLORECTAL CANCERS

Hierarchical clustering analysis revealed that memory CD8$^+$/γδ T cell phenotypes clustered in a tissue-specific manner (Figure 6.2A). Two CD8$^+$CD103$^+$PD-1$^+$ populations (#60 and 96), distinguished by CD161 expression, were present in tumour tissues (constituting up to 28.2% of CD45$^+$ cells) and infrequent in all other samples (Figure 6.2B-C), with the exception of one lymph node sample that was found to be infiltrated by tumour cells upon histological examination (data not shown). These CD8$^+$CD103$^+$PD-1$^+$ cells were further characterized by the co-expression of CD69, FAS, HLA-DR, and CD38 (Figure 6.2B). Interestingly, the CD161$^-$ counterpart of CD8$^+$CD103$^+$PD-1$^+$ T cells (#60) was particularly abundant in MMR-deficient tumours as compared to MMR-proficient tumours (Supplementary Figure 6.3). Within the CD8$^+$CD103$^+$PD-1$^+$CD38$^+$ subset, we observed co-expression of CD39 (Supplementary Figure 6.3), a marker that has recently been found to identify tumour-reactive CD8$^+$ T cells[22,23]. Next to these tumour-resident cells, a cluster (#61) with a similar phenotype but lacking HLA-DR, PD-1, FAS, and possessing a lower expression of CD38 was present in both tumour and healthy colorectal samples (Figure 6.2B-C), and may represent a non-activated counterpart. Single-cell RNA-sequencing revealed that CD8$^+$ T cells in colorectal tumours expressed cytolytic molecules (e.g. *GZMA*, *GZMB*, *GZMH, PRF1*) (Figure 6.2D). Furthermore, they displayed expression of the immune checkpoint molecule *LAG3* (Figure 6.2D).

Strikingly, a TCRγδ$^+$CD103$^+$PD-1$^+$ population (#99) was almost exclusively found in MMR-deficient tumours, constituting up to 8.4% of CD45$^+$ cells (Figure 6.2B-C). These γδ T cells had a phenotype similar to the CD8$^+$CD103$^+$PD-1$^+$ cells, as defined by co-expression of CD69, FAS, CD38, and HLA-DR (Figure 6.2B). An HLA-DR$^-$PD-1$^-$ counterpart of these cells (#97 and 101) was also observed in colorectal healthy mucosa and MMR-proficient tumours, and may represent a non-activated form of the CD103$^+$PD-1$^+$ γδ T cells in the tumour microenvironment (Figure 6.2B-C). We analyzed the cytotoxic potential of the tumour-resident γδ T cells by flow cytometry and determined that these were capable of producing IFN-γ and granzyme B/perforin upon stimulation with PMA/ionomycin (Supplementary Figure 6.4).

6.3.3 ICOS$^+$ AND ACTIVATED CD4$^+$ T CELLS ARE DOMINANT, TUMOUR TISSUE-SPECIFIC T CELL POPULATIONS IN BOTH MISMATCH REPAIR-DEFICIENT AND REPAIR-PROFICIENT COLORECTAL CANCERS

Next, we determined the cell surface phenotype of memory CD4$^+$ T cells in CRC patients. Memory CD4$^+$ T cells also distributed in a tissue-specific manner (Figure 6.3A). Here, a large population of CD4$^+$ICOS$^+$CD27$^-$ cells (#20 and 58) constituted up to 21.1% of CD45$^+$ cells in CRCs, while being absent in all other tissues with the exception of tumour-positive lymph node samples (Figure 6.3B-C). Part of this population co-expressed CD161 and PD-1 (#58), whereas the other part was negative for these markers but expressed high levels of CD25 (#20), indicative of a regulatory-like phenotype (Figure 6.3B). Flow cytometry analysis confirmed the expression of FOXP3 in 91-98% of ICOS$^+$ CD4$^+$CD45RO$^+$CD25$^+$CD127^{low} T cells in colorectal tumours (Supplementary Figure 6.5). Interestingly, the ICOS$^+$ CD4$^+$ T cells were present in MMR-deficient as well as MMR-proficient tumours to a similar extent (Figure 6.3B-C).

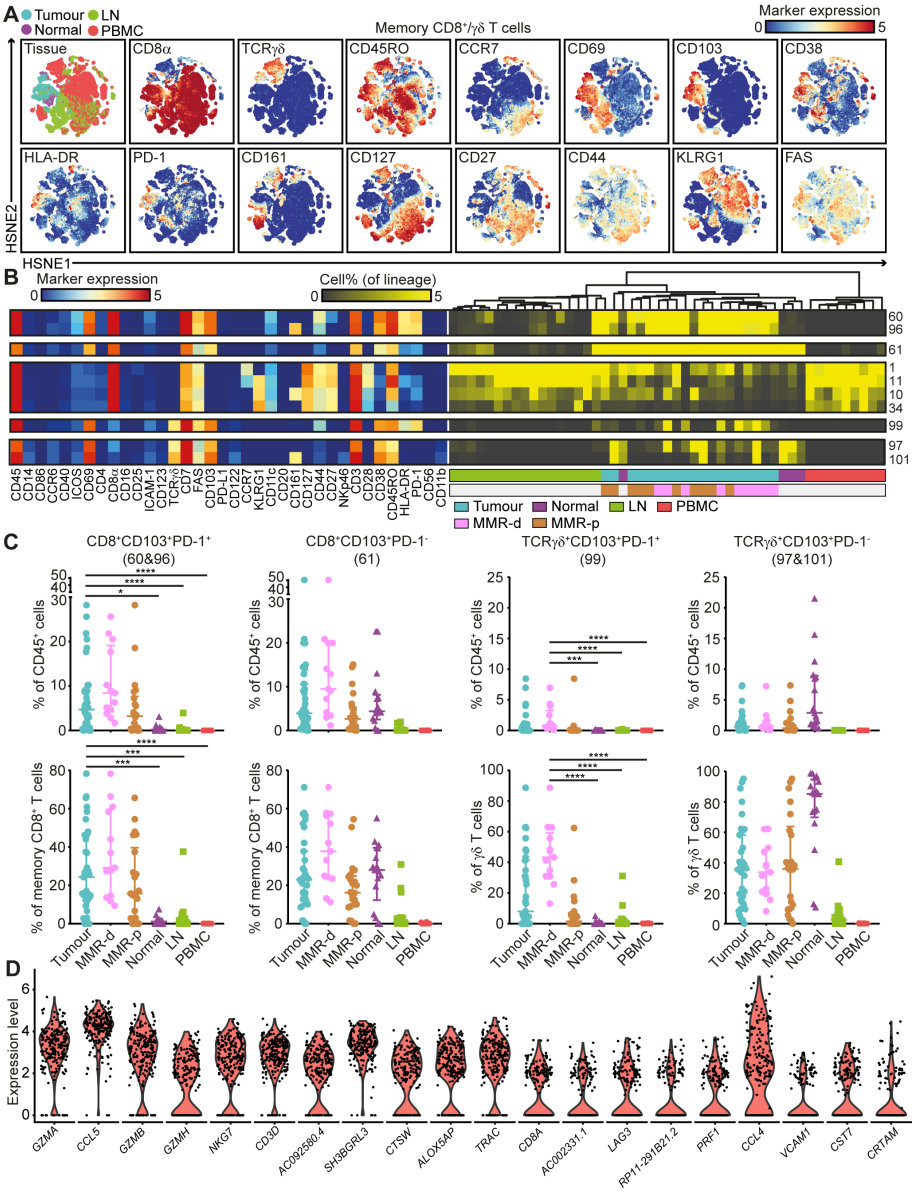


Figure 6.2 Activated CD8$^+$ and γδ T cells are tumour tissue-specific and enriched in mismatch repair-deficient colorectal cancers. (A) HSNE embedding of 1.6*10^5 landmarks representing the memory CD8$^+$/γδ T cell compartment (1.1*10^6 cells) from the discovery cohort of CRC patients colored by tissue type (first plot) and relative expression of indicated markers. **(B)** A heatmap showing median marker expression values (left) and a heatmap showing frequencies of selected memory CD8$^+$/γδ T cell clusters (right). Hierarchical clustering was performed on cluster frequencies using Spearman's rank correlation. Color bars indicate tissue type. **(C)** Frequencies of selected memory CD8$^+$/γδ T cell clusters among CRC patients' tissues (N=35, further subdivided into MMR-deficient (N=13) and MMR-proficient (N=22)), colorectal healthy mucosa (N=17), tumour-associated lymph nodes (N=26), and peripheral blood (N=19) as percentage of total CD45$^+$ cells (upper panel) and memory CD8$^+$ or γδ T cells (lower panel). Cluster IDs correspond to the ones in (B). Bars indicate median ± IQR. Each dot represents an individual sample. Data from 22 independent experiments with mass cytometry. *P<0.05, ***P<0.001,

****P<0.0001 by Kruskal-Wallis test with Dunn's test for multiple comparisons. **(D)** Violin plot showing log-transformed expression levels of the top 20 differentially expressed genes within $CD8^+$ T cells (N=217) analyzed by single-cell RNA-sequencing on $CD45^+$ cells from 7 tumours (Figure 6.1D). Each dot represents a single cell. LN; lymph node, MMR-d; mismatch repair-deficient, MMR-p; mismatch repair-proficient, PBMC; peripheral blood mononuclear cell.

In addition, $CD4^+CD103^+PD\text{-}1^+$ cells (#85 and 86), which constituted up to 23.8% of $CD45^+$ cells, were also enriched in tumour tissues (Figure 6.3B-C). Strikingly, several features of these cells mirrored our observations in the $CD8^+/\gamma\delta$ compartment, including a tissue-resident phenotype defined by co-expression of CD69, FAS, CD38, and HLA-DR (Figure 6.3B). Moreover, expression of CD161 also subdivided $CD4^+CD103^+PD\text{-}1^+$ T cells into a positive (#85) and negative (#86) population, where $CD161^-$ cells were more abundant in MMR-deficient as compared to MMR-proficient tumours (Supplementary Figure 6.3). In contrast to the tumour-resident $CD8^+$ and $\gamma\delta$ T cells, a non-activated counterpart could not be detected for these cells.

While $ICOS^+$ regulatory T cells (Tregs) were tumour tissue-specific, $ICOS^-CD25^+CD127^-$ Tregs (#13-73) were found in both tumour-associated lymph nodes and CRC tissues (Figure 6.3B-C). Lastly, immune cell populations such as $CD4^+CD27^+CD127^+$ central memory ($CCR7^+CD45RO^+$) cells (#1-37) were more abundant in peripheral blood and lymph nodes (Figure 6.3B-C). The expression of *ICOS* on $CD4^+$ T cells was confirmed by single-cell RNA-sequencing, which also revealed the expression of *TNFRSF4 (OX40R)* and *TNFRSF18 (GITR)* (Figure 6.3D). t-SNE analysis revealed the co-expression of all three immunotherapeutic targets by $CD4^+$ T cells (Supplementary Figure 6.6).

6.3.4 $CD127^-CD56^+CD45RO^+$ ILCS ARE THE PREVALENT ILC POPULATION IN MISMATCH REPAIR-DEFICIENT COLORECTAL TUMOURS

Mass cytometric profiles of the innate lymphoid compartment revealed the presence of three distinct Lin^-CD7^+ cell clusters: $CD127^-CD56^+CD45RO^-$ natural killer (NK) cells (90.4%), $CD127^+$ ILCs (3.4%), and a cluster of $CD127^-CD56^+CD45RO^+$ cells (6.2%) (Figure 6.4A). Analysis of cluster frequencies demonstrated that $CD56^{dim}CD16^{bright}$ NK cells (#33-4) were present in high frequencies in peripheral blood, whereas $CD56^{bright}CD16^{dim}$ NK cells (#10-82) were the dominant NK-type in lymph node samples (Figure 6.4B-C). $CD127^+$ ILCs (#6-9) were more abundant in healthy mucosa, lymph nodes and MMR-proficient tumours, and displayed a $KLRG1^-$ phenotype, characteristic of ILC3 cells (Figure 6.4B-C). Strikingly, the $CD127^-CD56^+CD45RO^+$ ILCs (#87,95,92,97) were enriched in tumour tissues, accounting for up to 80% of the innate lymphoid compartment (Figure 6.4B-C). Moreover, they were particularly abundant in MMR-deficient tumours, especially $CD161^-$ populations (#95, 92 and 97) (Figure 6.4B-C). The $CD127^-CD56^+CD45RO^+$ ILC population has recently been identified in human fetal intestine as intermediate-ILCs[24]. Consistent with that work, hierarchical clustering positioned the $CD127^-CD56^+CD45RO^+$ ILCs in between NK cells and $CD127^+$ ILCs (Figure 6.4B). We observed co-expression of CD69 and CD103 on all $CD127^-CD56^+CD45RO^+$ ILCs, but differential expression of CD16, ICAM-1, FAS, CD11c, CD161, CD44 and HLA-DR, indicative of further heterogeneity within this cell cluster (Figure 6.4B).

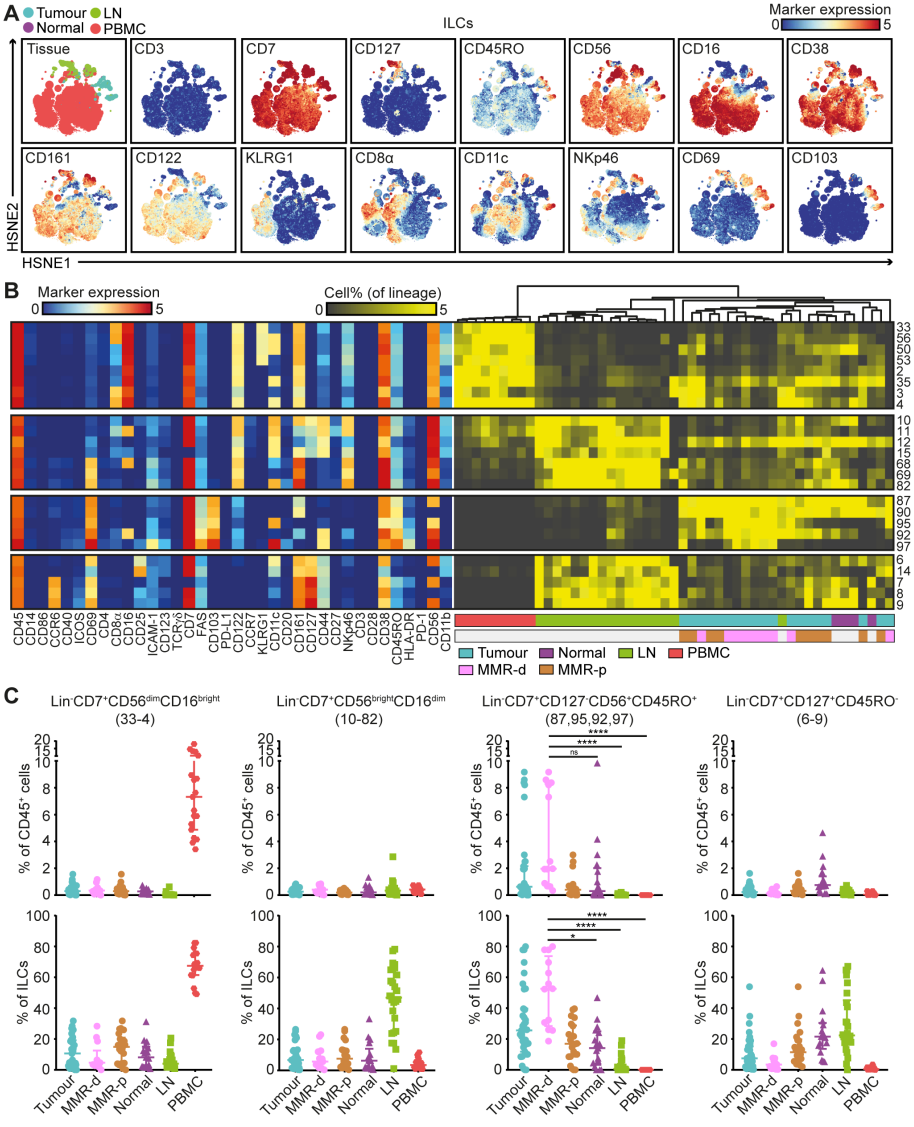


Figure 6.3 CD127$^-$CD56$^+$CD45RO$^+$ ILCs are the prevalent ILC population in mismatch repair-deficient colorectal tumours. (A) HSNE embedding of 5.5*10^4 landmarks representing the innate lymphoid compartment (0.4*10^6 cells) from the discovery cohort of CRC patients colored by tissue type (first plot) and relative expression of indicated markers. **(B)** A heatmap showing median marker expression values (left) and a heatmap showing frequencies of selected ILC clusters (right). Hierarchical clustering was performed on cluster frequencies using Spearman's rank correlation. Color bars indicate tissue type. **(C)** Frequencies of selected innate lymphoid clusters among CRC patients' tissues (N=35, further subdivided into MMR-deficient (N=13) and MMR-proficient (N=22)), colorectal healthy mucosa (N=17), tumour-associated lymph nodes (N=26), and peripheral blood (N=19) as percentage of total CD45$^+$ cells (upper panel) and ILCs (lower panel). Cluster IDs correspond to the ones in (B). Bars indicate median ± IQR. Each dot represents an individual sample. Data from 22 independent experiments with mass cytometry. NS, not significant, *P<0.05, ****P<0.0001 by Kruskal-Wallis test with Dunn's test for multiple comparisons. ILC; innate lymphoid cell, LN; lymph node, MMR-d; mismatch repair-deficient, MMR-p; mismatch repair-proficient, PBMC; peripheral blood mononuclear cell.

6.3.5 TUMOUR-RESIDENT ILCS ARE INVOLVED IN THE ANTI-TUMOUR IMMUNE RESPONSE

Single-cell RNA-sequencing unveiled high expression levels of cytotoxic molecules (e.g. *GNLY*, *PRF1*, *GZMA*, *GZMB*) in the ILC cluster (Figure 6.5A). In addition, we observed the presence of transcripts for a member of the killer-cell immunoglobulin-like receptor (KIR) family, *KIR2DL4* (Figure 6.5A). We performed additional single-cell RNA-sequencing on CD45$^+$ cells from one MMR-deficient tumour with high numbers of Lin$^-$CD7$^+$CD127$^-$CD56$^+$CD45RO$^+$ ILCs (70% of the ILC cluster), as revealed by mass cytometry data. Here, we also observed high expression levels of cytotoxic molecules (e.g. *GNLY*, *PRF1*, *GZMA*) as well as the expression of *KIR2DL4* and *KIR3DL2* in the ILC cluster (Supplementary Figure 6.7). Cell surface expression of KIRs was confirmed by flow cytometry in Lin$^-$CD7$^+$CD127$^-$CD56$^+$CD45RO$^+$ ILCs from this tumour (Supplementary Figure 6.7).

To further investigate functional properties of tumour-resident lymphocytes, we designed a flow cytometry antibody panel to analyze the cytotoxic potential of Lin$^-$CD7$^+$CD127$^-$CD56$^+$CD45RO$^+$ ILCs, Lin$^-$CD7$^+$CD127$^-$CD56$^+$CD45RA$^+$ NK cells, and memory CD8$^+$ T cells in CRC tissues. Strikingly, up to 82.3% of unstimulated CD127$^-$CD56$^+$CD45RO$^+$ ILCs displayed granzyme B/perforin expression in the tumour tissues (Figure 6.5B). Granzyme B/perforin expression by the ILCs was most abundant in MMR-deficient cancers as compared to MMR-proficient cancers (Figure 6.5C). Interestingly, the cytotoxic capacity of CD127$^-$CD56$^+$CD45RO$^+$ ILCs was accompanied by similar profiles in CD127$^-$CD56$^+$CD45RA$^+$ NK cells and memory CD8$^+$ T cells across samples (Figure 6.5C), suggesting a coordinated cytotoxic innate and adaptive immune response in CRC tissues.

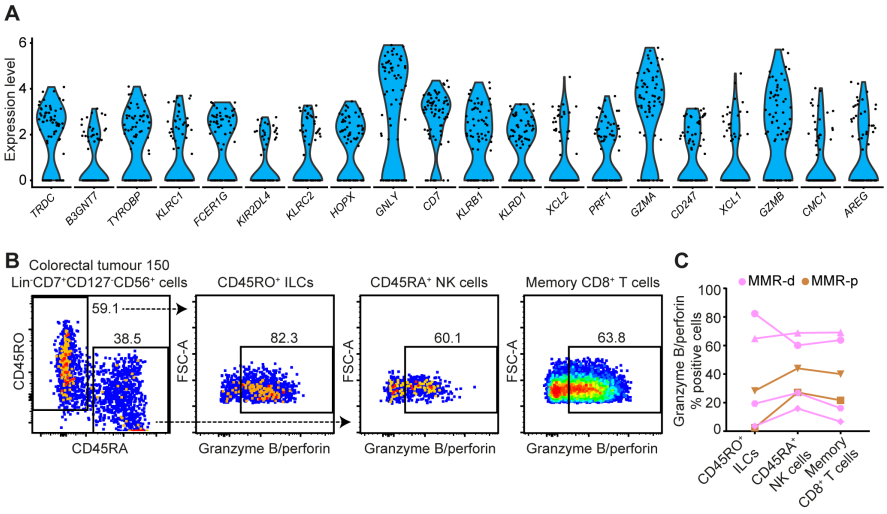


Figure 6.4 Tumour-resident ILCs are involved in the anti-tumour immune response. (A) Violin plot showing log-transformed expression levels of the top 20 differentially expressed genes within ILCs (N=74) analyzed by single-cell RNA-sequencing on CD45$^+$ cells from 7 tumours (Figure 6.1D). Each dot represents a single cell. **(B)** Representative plots of a MMR-deficient tumour sample analyzed by flow cytometry without stimulation showing the distinction between CD45RO$^+$ ILCs and CD45RA$^+$ NK cells within Lin$^-$CD7$^+$CD127$^-$CD56$^+$ cells (first plot), and their expression of cytotoxic molecules. **(C)** Granzyme B/perforin expression in different immune cell populations of CRC tissues (N=6, of which 4 MMR-deficient and 2 MMR-proficient). Dot shape indicates similar tumour samples. Data from three independent experiments with flow cytometry. ILC; innate lymphoid cell, MMR-d; mismatch repair-deficient, MMR-p; mismatch repair-proficient, NK; natural killer.

To investigate the spatial localization of the ILCs in CRCs, we applied 6-colour multispectral immunofluorescence to frozen tissue sections of 4 MMR-deficient and 4 MMR-proficient CRCs. We simultaneously detected CD3, TCRαβ, CD127, CD7, CD45RO, and DAPI. We identified $CD3^-TCR\alpha\beta^-CD127^-CD7^+CD45RO^+$ ILCs in the tumours (Figure 6.6A-B), and observed an increased presence of these cells in MMR-deficient as compared to MMR-proficient CRCs (Figure 6.6C). Interestingly, the $CD3^-TCR\alpha\beta^-CD127^-CD7^+CD45RO^+$ ILCs frequently displayed an intraepithelial localization in agreement with their $CD103^+CD69^+$ tissue-resident phenotype (Figure 6.6A).

6.3.6 IMMUNE-SYSTEM-WIDE ANALYSIS REVEALS CORRELATIONS BETWEEN INNATE AND ADAPTIVE IMMUNE CELL SUBSETS IN COLORECTAL CANCER

Lastly, we integrated the identified immune cell clusters across all major immune lineages (N=218) in one immune-system-wide analysis to characterize the samples according to tissue type, MMR status, and available clinico-pathological parameters. The integrated t-SNE analysis confirmed the unique immune composition in the different tissue types, and visualized the top ten ranked immune cell clusters contributing to the distinctive clustering patterns of the samples (Supplementary Figure 6.8). No association was observed with clinical stage while differences related to tumour location and HLA class I expression can be attributed to features that distinguish MMR-deficient and -proficient CRCs (Supplementary Figure 6.8).

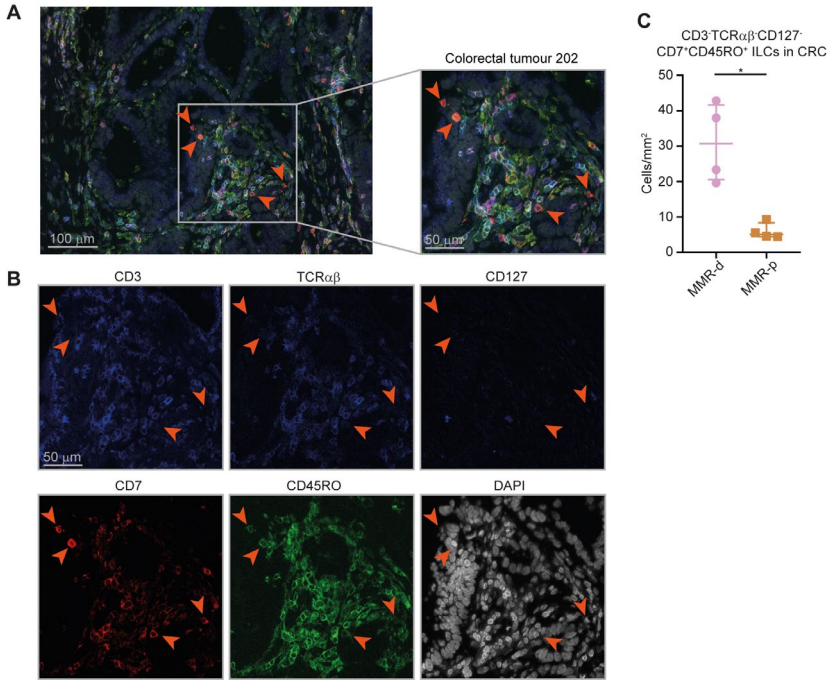


Figure 6.5 Higher cell density of $CD127^-CD45RO^+$ ILCs in mismatch repair-deficient colorectal cancers. (A and B) Representative image of the immunofluorescence microscopic detection of $CD3^-TCR\alpha\beta^-CD127^-CD7^+CD45RO^+$ ILCs in a MMR-deficient tumour, showing CD3 (colored in blue), TCRαβ (colored in blue), CD127 (colored in blue), CD7 (colored in red), CD45RO (colored in green), and DAPI (colored in grey) as nuclear counterstain. **(C)** Frequencies of $CD3^-TCR\alpha\beta^-CD127^-CD7^+CD45RO^+$ ILCs in 4 MMR-deficient and 4 MMR-proficient CRCs. *P<0.05 by Mann-Whitney U test. ILC; innate lymphoid cell, MMR-d; mismatch repair-deficient, MMR-p; mismatch repair-proficient.

Spearman's rank correlation analysis performed on the top ten ranked unique immune cell clusters of each tissue type revealed strong correlations between the presence of $CD127^-$ $CD56^+CD45RO^+$ ILCs (ILC97,92,95) and the presence of $CD103^+PD\text{-}1^+$ cytotoxic (CD8memory60,96), helper (CD4memory85,86), and γδ (TCRγδ99) T cell populations in MMR-deficient CRCs (Figure 6.7, Supplementary Table 6.5). In contrast, MMR-proficient tumours were characterized by the presence of several myeloid populations (Figure 6.7, Supplementary Table 6.5).

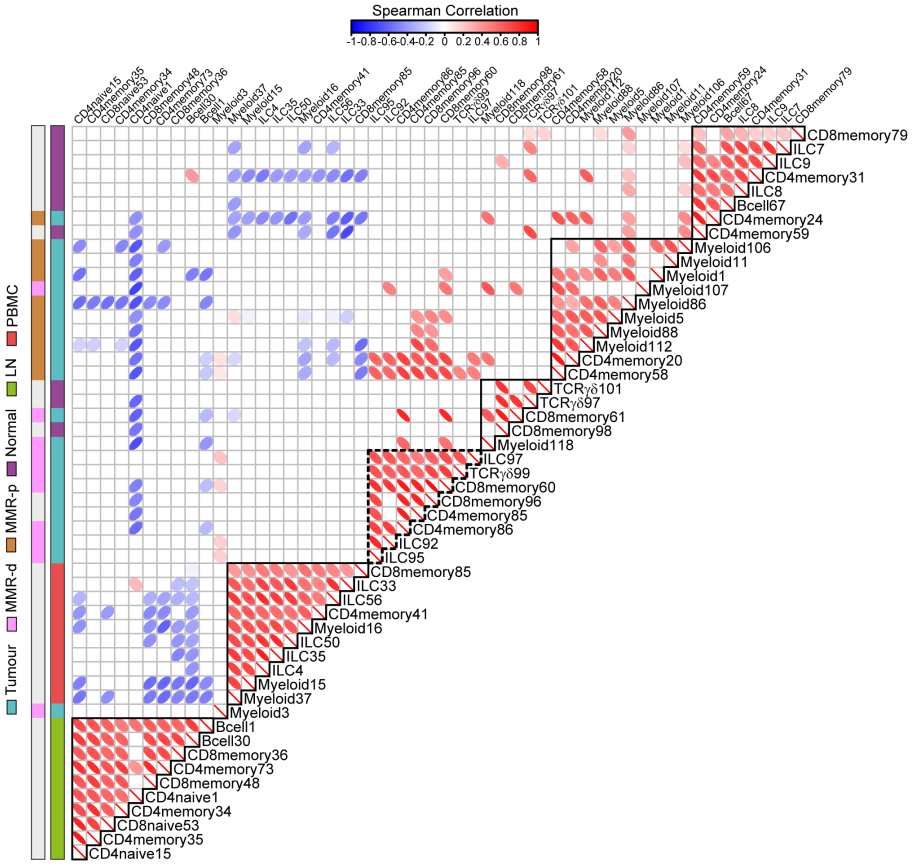


Figure 6.6 Immune-system-wide analysis reveals correlations between innate and adaptive immune cell subsets in colorectal cancer. Matrix showing correlations (Spearman's ρ, see Supplementary Methods) between unique top ten ranked immune cell clusters for each tissue type (shown in Supplementary Figure 6.8) based on cell percentages (of total $CD45^+$ cells) corresponding to 97 samples from 31 CRC patients. Color and shape of the ellipses in the heatmap indicate the strength of the correlation. Only significant correlation coefficients are shown. Color bars indicate tissue type. Coefficient and *P*-values of correlations for CRC tissues are shown in Supplementary Table 6.5. *P*-values were adjusted for multiple testing using the Benjamini-Hochberg procedure. Data from 22 independent experiments with mass cytometry. ILC; innate lymphoid cell, LN; lymph node, MMR-d; mismatch repair-deficient, MMR-p; mismatch repair-proficient, PBMC; peripheral blood mononuclear cell.

6.4 DISCUSSION

We applied mass cytometry to comprehensively analyze the immune landscape of CRCs at single-cell level in tumour and healthy tissues. Our analysis revealed tumour tissue-specific immune signatures across the innate and adaptive immune compartments of CRC. Immunohistochemistry, flow cytometry, and recent transcriptomic approaches have provided insight into the complexity of tumour immune landscapes[25–29]. However, the number of markers that can be simultaneously assayed in immunohistochemistry or flow cytometry is limited, and bulk transcriptomic studies do not allow for discrimination of phenotypes at the cellular level[30,31]. In mass cytometry over 40 markers can be simultaneously analyzed at single-cell level, providing a unique opportunity to obtain a comprehensive overview of tumour-resident lymphocytes[32,33]. Here, we combined mass cytometry phenotypes with functional, transcriptional, and spatial analyses of tumour-resident immune cell populations in CRC.

Within the innate compartment, we observed that a previously unappreciated innate lymphoid population, $Lin^-CD7^+CD127^-CD56^+CD45RO^+$ ILCs, is enriched in MMR-deficient tumours and displayed cytotoxic activity. *In-situ* detection of the ILCs confirmed a higher cell density in MMR-deficient CRCs, and showed a frequent intraepithelial localization. This is in line with their tissue-resident phenotype ($CD103^+CD69^+$), and supports an active role for these cells in the anti-tumour immune response. The ILCs resemble previous descriptions of TCR^-CD103^+ cells in mice that were found to express granzyme B[34]. Additionally, a unique subset of NK cells has been found in several human tissues and was described as $NKp44^+CD103^+$ intraepithelial ILC1-like[35,36]. In contrast to $NKp44^+CD103^+$ ILC1, the $CD127^-CD56^+CD45RO^+$ ILCs identified here lacked CD122 and NKp46 expression (Figure 6.4B), and showed low levels of NKp44 (data not shown). These variable marker expression patterns most likely represent additional levels of plasticity and heterogeneity within ILC subsets. Single-cell RNA-sequencing revealed the presence of transcripts for *KIR2DL4* and *KIR3DL2* in the ILC cluster, which hints towards potential activation mechanisms[37]. Common ligands of KIRs include HLA class I molecules[38,39], and loss of HLA class I expression has been described to occur in the majority of MMR-deficient CRCs[40–42]. It is tempting to speculate that $CD127^-CD56^+CD45RO^+$ ILC-mediated cytotoxicity towards such HLA-loss variants may contribute to the anti-tumour response in MMR-deficient CRCs, a link that requires further investigation.

The presence of $CD127^-CD56^+CD45RO^+$ ILCs strongly correlated with tissue-resident $CD103^+CD69^+$ $\gamma\delta$ T cells co-expressing activation markers HLA-DR, CD38, and PD-1 in MMR-deficient CRCs. It has been shown that human peripheral blood $\gamma\delta$ T cells can express PD-1 and exhibit natural killer-like activity[43]. The expression of PD-1, in conjunction with their cytotoxic potential, suggest an active role of tumour-resident $\gamma\delta$ T cells in the anti-tumour immune response and potentially as targets for PD-1 checkpoint blockade. This will be subject of further studies.

Within the adaptive compartment, we found dominant, tumour tissue-specific $CD8^+$ and $CD4^+$ T cell populations that displayed a highly similar activated tissue-resident phenotype. Such $CD8^+$ T cell populations have been described in ovarian cancer[44,45], lung cancer[46], and recently in melanoma[47], cervical carcinoma[48] and CRC[22,23], and their presence was associated with an improved clinical prognosis. Single-cell RNA-sequencing revealed that $CD8^+$ T cells in colorectal tumours showed a cytotoxic profile, indicative of potential anti-tumour reactivity. In addition, we found a dominant tumour tissue-specific population of $ICOS^+$ $CD4^+$ T cells. ICOS belongs to the CD28/CTLA-4 family and serves as a co-stimulatory molecule for T cell activation[49]. Activation of ICOS by agonists has been proposed for anti-cancer treatment[50].

Here, we identified a CD161$^+$PD-1$^+$ as well as a CD25$^+$ population of tumour-resident ICOS$^+$ CD4$^+$ T cells. The latter corresponds to a regulatory T cell subset displaying high levels of FOXP3 expression, that, interestingly, expressed higher levels of ICOS as compared to the CD161$^+$PD-1$^+$ counterpart. The use of ICOS agonists may, therefore, also result in activation of ICOS$^+$ T cells with suppressive and regulatory properties in the tumour microenvironment. In contrast to the tumour-resident CD8$^+$ T cells, ICOS$^+$ CD4$^+$ T cells were present in both MMR-deficient and MMR-proficient tumours to a similar extent.

We observed CD161$^+$ and CD161$^-$ counterparts of tumour-resident cytotoxic and helper T cells, and CD127$^-$CD56$^+$CD45RO$^+$ ILCs. CD161 has been shown to mark a subset of tissue-resident memory CD8$^+$ T cells with enhanced effector function and cytokine production[51,52]. In our study, the CD161$^-$ counterpart of the tumour-resident T cell and ILC populations was particularly enriched in MMR-deficient CRCs as compared to MMR-proficient CRCs. The functional relevance of this observation will be subject of future studies. Nevertheless, we observed increased CD161 expression in PD-1 high cells as compared to PD-1 intermediate/negative cells for tumour-resident CD8$^+$ and CD4$^+$ T cell populations (Supplementary Figure 6.9). As PD-1 high cells in human cancer have been associated with a state of T cell dysfunction[53–55], CD161 expression could be an additional marker for this functional state.

Interestingly, we identified what could be the non-activated counterparts of the CD103$^+$PD-1$^+$ cytotoxic and $\gamma\delta$ T cells in both tumour and healthy colorectal tissues. Mobilization and activation of these cells from the colorectal healthy mucosa to the tumour tissue may be beneficial for immunotherapy in CRC. Strikingly, while lymph nodes are traditionally viewed as key players of anti-tumour immune responses, we did not detect non-activated precursors of tumour-resident immune cell populations in the lymph node samples, with the exception of tumour-positive lymph nodes. Furthermore, we observed that lymph nodes harbored a large population of CD4$^+$CD25$^+$CD127$^-$ Tregs, suggesting they might be a primary source of Tregs in the cancer microenvironment. The tumour-resident immune cell populations were also not mirrored in peripheral blood, although the in-depth investigation of their presence in these tissues with complementary approaches should be conducted.

It should be noted that the mass cytometry antibody panel was primarily developed to characterize T cell, $\gamma\delta$ T cell and ILC compartments, and in future studies additional efforts are required to further explore the myeloid and B cell compartment. Furthermore, the number and pattern of infiltrating lymphocytes can be influenced by various tumour characteristics. In this study we have shown profound differences in lymphocytic infiltration that distinguish MMR-deficient from MMR-proficient CRCs. Other factors not investigated in this study that can influence the infiltration of lymphocytes in tumours include for instance occurrence of somatic mutations (neoantigens) and the co-occurrence of inflammatory bowel disease. Although the results are of preliminary nature, they point to the involvement of additional subsets than T cells in immune responses to CRC, particularly ILCs and $\gamma\delta$ T cells. This is especially relevant in the context of responses to checkpoint blockade therapy in absence of HLA class I expression[56]. Future approaches might opt for an in-depth investigation of these specific lineages for a detailed characterization of phenotypes that complement the markers used in this study. The next step will be to investigate the involvement of these subsets in the clinical setting of patients treated by checkpoint blockade.

In conclusion, we identified a previously unappreciated innate immune cell population that was specifically enriched in CRC tissues, displayed cytotoxic activity, and strongly contributed to a data-driven distinction between immunogenic (MMR-deficient) and non-immunogenic

(MMR-proficient) tumours. Furthermore, we revealed strong correlations between the presence of these innate cells and tumour-resident $CD8^+$, $CD4^+$, and $\gamma\delta$ T cells with an activated phenotype in MMR-deficient tumours that together may play a critical role in tumour control.

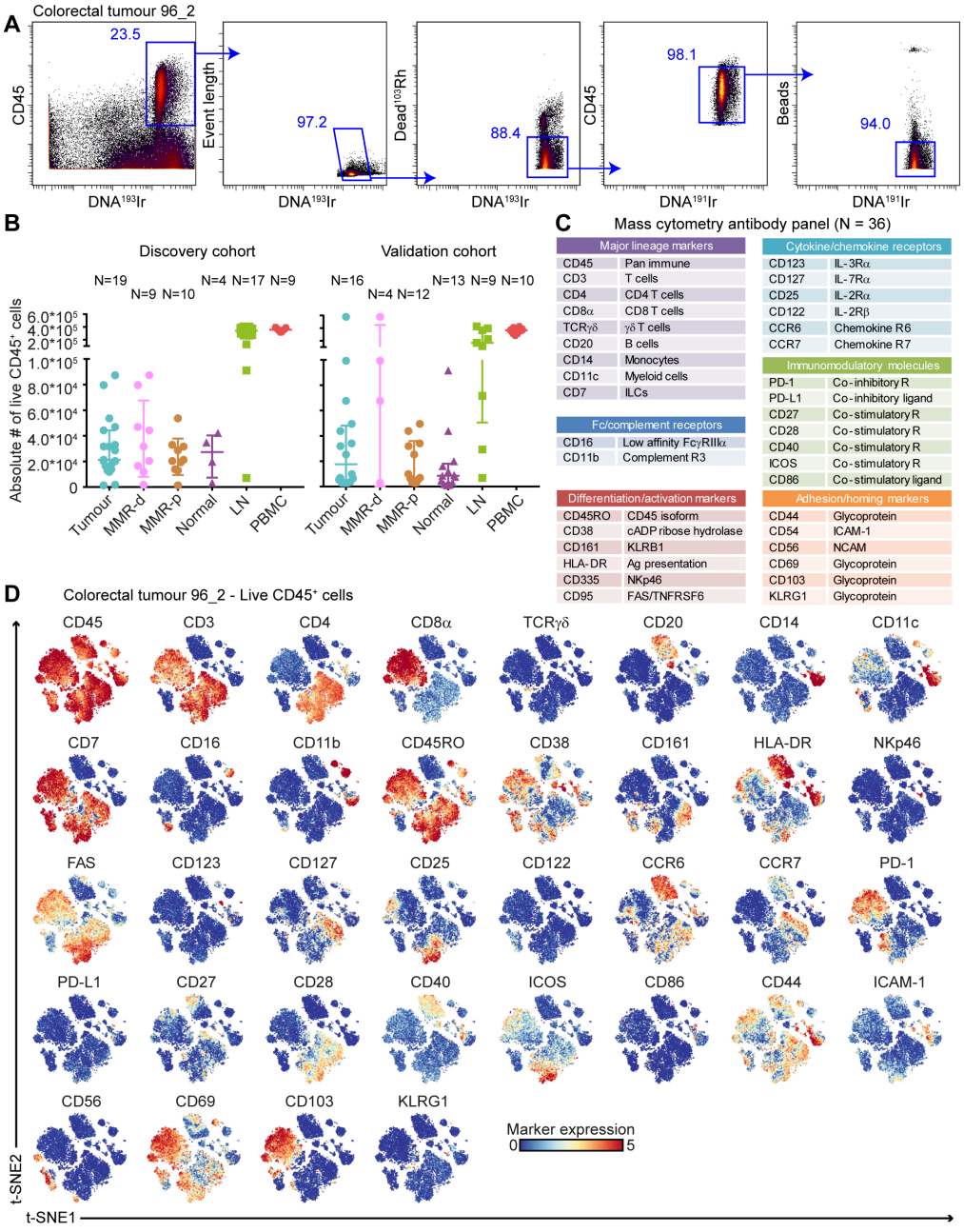
BIBLIOGRAPHY

1. Hodi, F. S. *et al.* Improved Survival with Ipilimumab in Patients with Metastatic Melanoma. *N. Engl. J. Med.* **363,** 711–723 (2010).
2. Topalian, S. L. *et al.* Safety, Activity, and Immune Correlates of Anti–PD-1 Antibody in Cancer. *N. Engl. J. Med.* **366,** 2443–2454 (2012).
3. Rizvi, N. A. *et al.* Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science (80-.).* **348,** 124–128 (2015).
4. Kelderman, S., Schumacher, T. N. & Kvistborg, P. Mismatch Repair-Deficient Cancers Are Targets for Anti-PD-1 Therapy. *Cancer Cell* **28,** 11–13 (2015).
5. Le, D. T. *et al.* Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science (80-.).* **357,** 409–413 (2017).
6. Ionov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D. & Perucho, M. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature* **363,** 558–561 (1993).
7. Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
8. Smyrk, T. C., Watson, P., Kaul, K. & Lynch, H. T. Tumor-infiltrating lymphocytes are a marker for microsatellite instability in colorectal carcinoma. *Cancer* **91,** 2417–2422 (2001).
9. De Miranda, N. F. C. C. *et al.* Infiltration of lynch colorectal cancers by activated immune cells associates with early staging of the primary tumor and absence of lymph node metastases. *Clin. Cancer Res.* **18,** 1237–1245 (2012).
10. Chevrier, S. *et al.* An Immune Atlas of Clear Cell Renal Cell Carcinoma. *Cell* **169,** 736–749 (2017).
11. Zheng, C. *et al.* Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. *Cell* **169,** 1342–1356.e16 (2017).
12. Kotecha, N., Krutzik, P. O. & Irish, J. M. Web-based analysis and publication of flow cytometry experiments. *Current Protocols in Cytometry* (2010). doi:10.1002/0471142956.cy1017s53
13. Höllt, T. *et al.* Cytosplore : Interactive Immune Cell Phenotyping for Large Single-Cell Datasets. in *Computer Graphics Forum (Proceedings of EuroVis 2016)* **35,** (2016).
14. Pezzotti, N., Höllt, T., Lelieveldt, B., Eisemann, E. & Vilanova, A. Hierarchical Stochastic Neighbor Embedding. in *Computer Graphics Forum (Proceedings of EuroVis 2016)* **35,** (2016).
15. Van Unen, V. *et al.* Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. *Nat. Commun.* **8,** 1–10 (2017).
16. Abdelaal, T. *et al.* Predicting Cell Populations in Single Cell Mass Cytometry Data. *Cytom. Part A* **95,** (2019).
17. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127 (2007).
18. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36,** 411–420 (2018).
19. Waltman, L. & Van Eck, N. J. A smart local moving algorithm for large-scale modularity-based community detection. *Eur. Phys. J. B* **86,** (2013).
20. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn.* **9,** 2579–2605 (2008).

21. Ijsselsteijn, M. E. *et al.* Cancer immunophenotyping by seven-colour multispectral imaging without tyramide signal amplification. *J. Pathol. Clin. Res.* **5,** 3–11 (2019).
22. Duhen, T. *et al.* Co-expression of CD39 and CD103 identifies tumor-reactive CD8 T cells in human solid tumors. *Nat. Commun.* **9,** (2018).
23. Simoni, Y. *et al.* Bystander CD8+ T cells are abundant and phenotypically distinct in human tumour infiltrates. *Nature* **557,** 575–579 (2018).
24. Li, N. *et al.* Mass cytometry reveals innate lymphoid cell differentiation pathways in the human fetal intestine. *J. Exp. Med.* **215,** 1383–1396 (2018).
25. Menon, A. G. *et al.* Immune system and prognosis in colorectal cancer: A detailed immunohistochemical analysis. *Lab. Investig.* **84,** 493–501 (2004).
26. Galon, J. *et al.* Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science (80-.).* **313,** 1960–1964 (2006).
27. Angelova, M. *et al.* Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol.* **16,** (2015).
28. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21,** 1350–1356 (2015).
29. Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **49,** 708–718 (2017).
30. Peterson, V. M. *et al.* Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* **35,** 936–939 (2017).
31. Schulz, D. *et al.* Simultaneous Multiplexed Imaging of mRNA and Proteins with Subcellular Resolution in Breast Cancer Tissue Samples by Mass Cytometry. *Cell Syst.* **6,** 25–36.e5 (2018).
32. Bandura, D. R. *et al.* Mass cytometry: Technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* **81,** 6813–6822 (2009).
33. Bendall, S. C., Nolan, G. P., Roederer, M. & Chattopadhyay, P. K. A deep profiler ’s guide to cytometry. *Trends Immunol.* **33,** 323–332 (2012).
34. Dadi, S. *et al.* Cancer Immunosurveillance by Tissue-Resident Innate Lymphoid Cells and Innate-like T Cells. *Cell* **164,** 365–377 (2016).
35. Fuchs, A. *et al.* Intraepithelial type 1 innate lymphoid cells are a unique subset of il-12- and il-15-responsive ifn-γ-producing cells. *Immunity* **38,** 769–781 (2013).
36. Simoni, Y. *et al.* Human Innate Lymphoid Cell Subsets Possess Tissue-Type Based Heterogeneity in Phenotype and Frequency. *Immunity* **46,** 148–161 (2017).
37. Wagtmann, N. *et al.* Molecular clones of the p58 NK cell receptor reveal immunoglobulin-related molecules with diversity in both the extra- and intracellular domains. *Immunity* **2,** 439–449 (1995).
38. Rajagopalan, S. & Long, E. O. KIR2DL4 (CD158d): An activation receptor for HLA-G. *Front. Immunol.* **3,** (2012).
39. Moretta, A. *et al.* P58 molecules as putative receptors for major histocompatibility complex (MHC) class I molecules in human natural killer (NK) cells. Anti-p58 antibodies reconstitute lysis of MHC class I-protected cells in NK clones displaying different specificities. *J. Exp. Med.* **178,** 597–604 (1993).
40. Dierssen, J. W. F. *et al.* HNPCC versus sporadic microsatellite-unstable colon cancers follow different routes toward loss of HLA class I expression. *BMC Cancer* **7,** (2007).
41. Kloor, M. *et al.* Immunoselective pressure and human leukocyte antigen class I antigen machinery defects in microsatellite unstable colorectal cancers. *Cancer Res.* **65,** 6418–6424 (2005).
42. Ijsselsteijn, M. E. *et al.* Revisiting immune escape in colorectal cancer in the era of immunotherapy. *Br. J. Cancer* **120,** 815–818 (2019).
43. Iwasaki, M. *et al.* Expression and function of PD-1 in human γδ T cells that recognize phosphoantigens. *Eur. J. Immunol.* **41,** 345–355 (2011).
44. Webb, J. R., Milne, K., Watson, P., DeLeeuw, R. J. & Nelson, B. H. Tumor-infiltrating

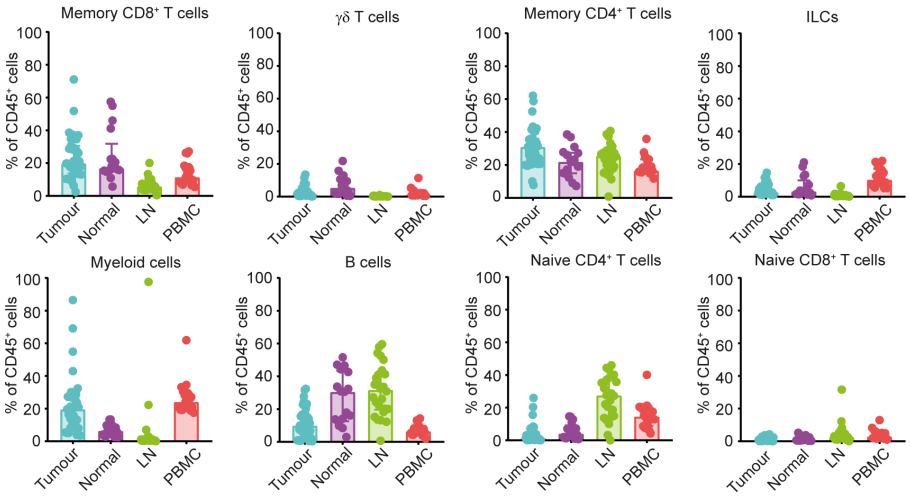
lymphocytes expressing the tissue resident memory marker cd103 are associated with increased survival in high-grade serous ovarian cancer. *Clin. Cancer Res.* **20,** 434–444 (2014).
45. Webb, J. R., Milne, K. & Nelson, B. H. PD-1 and CD103 are widely coexpressed on prognostically favorable intraepithelial CD8 T cells in human ovarian cancer. *Cancer Immunol. Res.* **3,** 926–935 (2015).
46. Djenidi, F. *et al.* CD8 + CD103 + Tumor–Infiltrating Lymphocytes Are Tumor-Specific Tissue-Resident Memory T Cells and a Prognostic Factor for Survival in Lung Cancer Patients . *J. Immunol.* **194,** 3475–3486 (2015).
47. Edwards, J. *et al.* CD103+ tumor-resident CD8+ T cells are associated with improved survival in immunotherapy-naïve melanoma patients and expand significantly during anti-PD-1 treatment. *Clin. Cancer Res.* **24,** 3036–3045 (2018).
48. Santegoets, S. J. *et al.* The anatomical location shapes the immune infiltrate in tumors of same etiology and affects survival. *Clin. Cancer Res.* **25,** 240–252 (2019).
49. Hutloff, A. *et al.* ICOS is an inducible T-cell co-stimulator structurally and functionally related to CD28. *Nature* **397,** 263–266 (1999).
50. Burris, H. A. *et al.* Phase 1 safety of ICOS agonist antibody JTX-2011 alone and with nivolumab (nivo) in advanced solid tumors; predicted vs observed pharmacokinetics (PK) in ICONIC. *J. Clin. Oncol.* **35,** 3033–3033 (2017).
51. Fergusson, J. R. *et al.* CD161 defines a transcriptional and functional phenotype across distinct human T cell lineages. *Cell Rep.* **9,** 1075–1088 (2014).
52. Fergusson, J. R. *et al.* CD161int CD8+ T cells: A novel population of highly functional, memory CD8+ T cells enriched within the gut. *Mucosal Immunol.* **9,** 401–413 (2016).
53. Thommen, D. S. *et al.* A transcriptionally and functionally distinct pd-1 + cd8 + t cell pool with predictive potential in non-small-cell lung cancer treated with pd-1 blockade. *Nat. Med.* **24,** (2018).
54. Kansy, B. A. *et al.* PD-1 status in CD8+ T cells associates with survival and anti-PD-1 therapeutic outcomes in head and neck cancer. *Cancer Res.* **77,** 6353–6364 (2017).
55. Kim, H. D. *et al.* Association Between Expression Level of PD1 by Tumor-Infiltrating CD8+ T Cells and Features of Hepatocellular Carcinoma. *Gastroenterology* **155,** 1936–1950.e17 (2018).
56. Middha, S. *et al.* Majority of B2M-Mutant and -Deficient Colorectal Carcinomas Achieve Clinical Benefit From Immune Checkpoint Inhibitor Therapy and Are Microsatellite Instability-High. *JCO Precis. Oncol.* 1–14 (2019). doi:10.1200/PO.18.00321

SUPPLEMENTARY MATERIALS

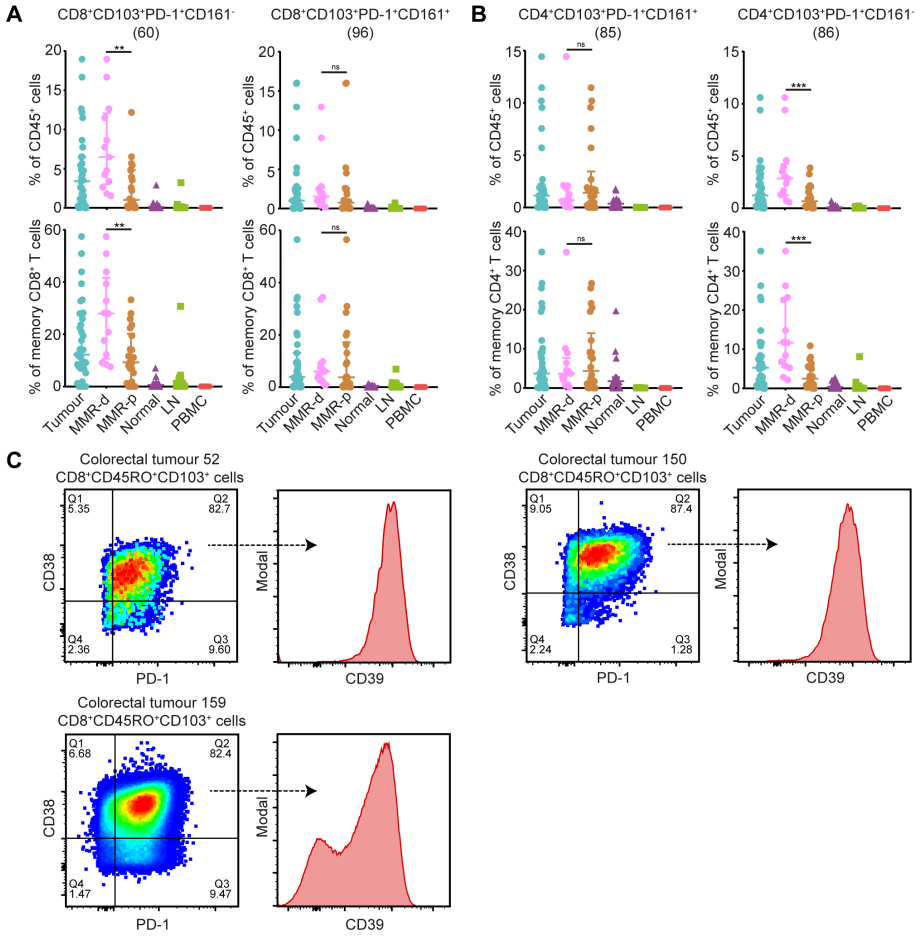


Supplementary Figure 6.1 Mass cytometry gating strategy and antibody expression patterns. **(A)** Mass cytometry gating strategy for single, live CD45⁺ cells of a representative colorectal tumour sample showing sequential gates with percentages. **(B)** Absolute number of live CD45⁺ cells of CRC tissues, colorectal healthy mucosa, tumour-associated lymph nodes, and peripheral blood samples of the discovery and validation cohort of CRC patients. Bars indicate median ± IQR. Each dot represents an

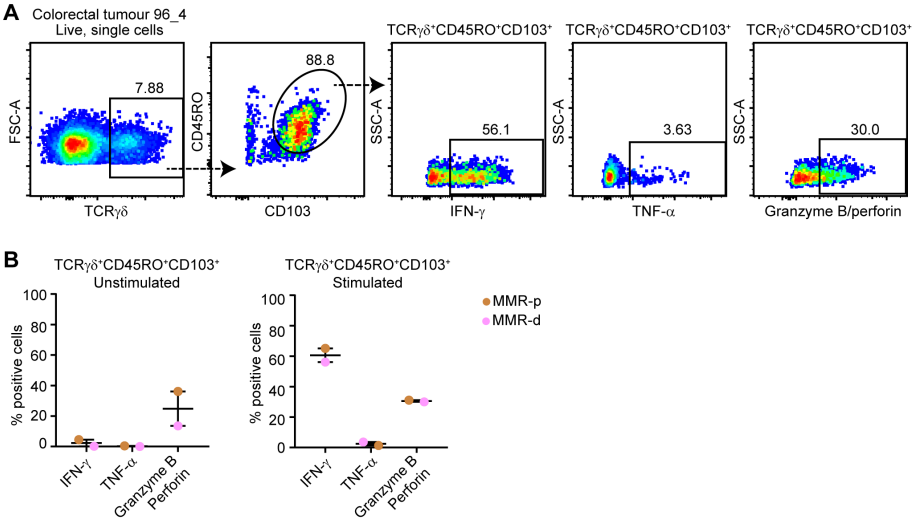
individual sample. Data from 22 independent experiments with mass cytometry. **(C)** Markers used to characterize immune cell phenotypes by mass cytometry. **(D)** t-SNE embedding showing marker expression patterns of each antibody on single, live CD45$^+$ cells (2.0*10^4) from the same tumour sample as shown in (A). Each dot represents a single cell. All markers are shown with an expression range of 0-5, with the exception of CD86 (0-3) due to lower sensitivity of the metal (^{115}In). CRC; colorectal cancer, LN; lymph node, MMR-d; mismatch repair-deficient, MMR-p; mismatch repair-proficient, PBMC; peripheral blood mononuclear cell, t-SNE; t-distributed stochastic neighbour embedding.



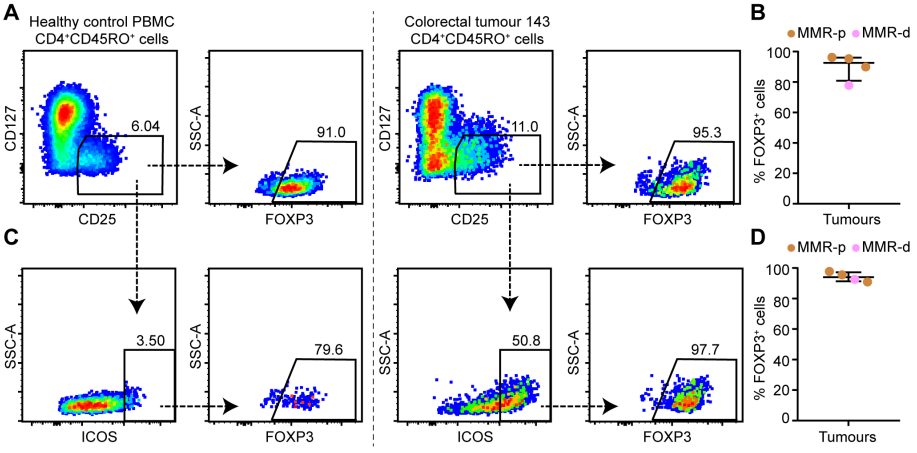
Supplementary Figure 6.2 Major immune lineage frequencies in different tissue types of colorectal cancer patients. Frequencies of major immune lineages across CRC tissues (N=35), colorectal healthy mucosa (N=17), tumour-associated lymph nodes (N=26), and peripheral blood (N=19) as percentage of total CD45$^+$ cells. Bars indicate median ± IQR. Each dot represents an individual sample. Data from 22 independent experiments with mass cytometry. ILC; innate lymphoid cell, LN; lymph node, PBMC; peripheral blood mononuclear cell.



Supplementary Figure 6.3 Characterization of tumour tissue-specific immune cell clusters corresponding to Figure 6.2 and 6.3. (A and B) Frequencies of $CD103^+PD\text{-}1^+CD161^-$ and $CD103^+PD\text{-}1^+CD161^+$ memory $CD8^+$ T cells **(A)** and $CD4^+$ T cells **(B)** among CRC tissues (N=35, further subdivided into MMR-deficient (N=13) and MMR-proficient (N=22)), colorectal healthy mucosa (N=17), tumour-associated lymph nodes (N=26), and peripheral blood (N=19) as percentage of total $CD45^+$ cells (upper panel) and memory $CD8^+$ or $CD4^+$ T cells (lower panel). Cluster IDs correspond to the ones in Figure 6.2B and 6.3B. Bars indicate median ± IQR. Each dot represents an individual sample. Data from 22 independent experiments with mass cytometry. NS, not significant, $^{**}P<0.01$, $^{***}P<0.001$ by Mann-Whitney U test. **(C)** Flow cytometry plots of colorectal tumours (N=3) showing the expression of CD39 within $CD8^+CD45RO^+CD103^+PD\text{-}1^+CD38^+$ cells. LN; lymph node, MMR-d; mismatch repair-deficient, MMR-p; mismatch repair-proficient, PBMC; peripheral blood mononuclear cell.

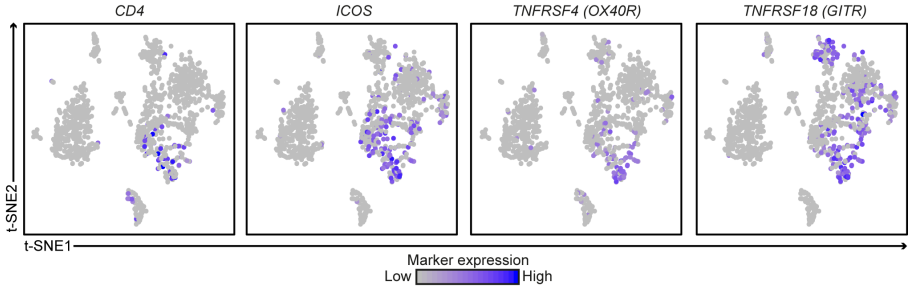


Supplementary Figure 6.4 Tumour-resident γδ T cells are capable of producing cytokines and cytotoxic molecules upon stimulation. (A) Flow cytometry plots of a MMR-deficient tumour sample showing the expression of cytokines and cytotoxic molecules by TCRγδ$^+$CD45RO$^+$CD103$^+$ cells upon stimulation with PMA/ionomycin. **(B)** IFN-γ, TNF-α, and granzyme B/perforin expression by TCRγδ$^+$CD45RO$^+$CD103$^+$ cells from a MMR-deficient and MMR-proficient CRC with and without stimulation with PMA/ionomycin. Bars indicate median ± IQR. Each dot represents an individual sample. MMR-d; mismatch repair-deficient, MMR-p; mismatch repair-proficient.

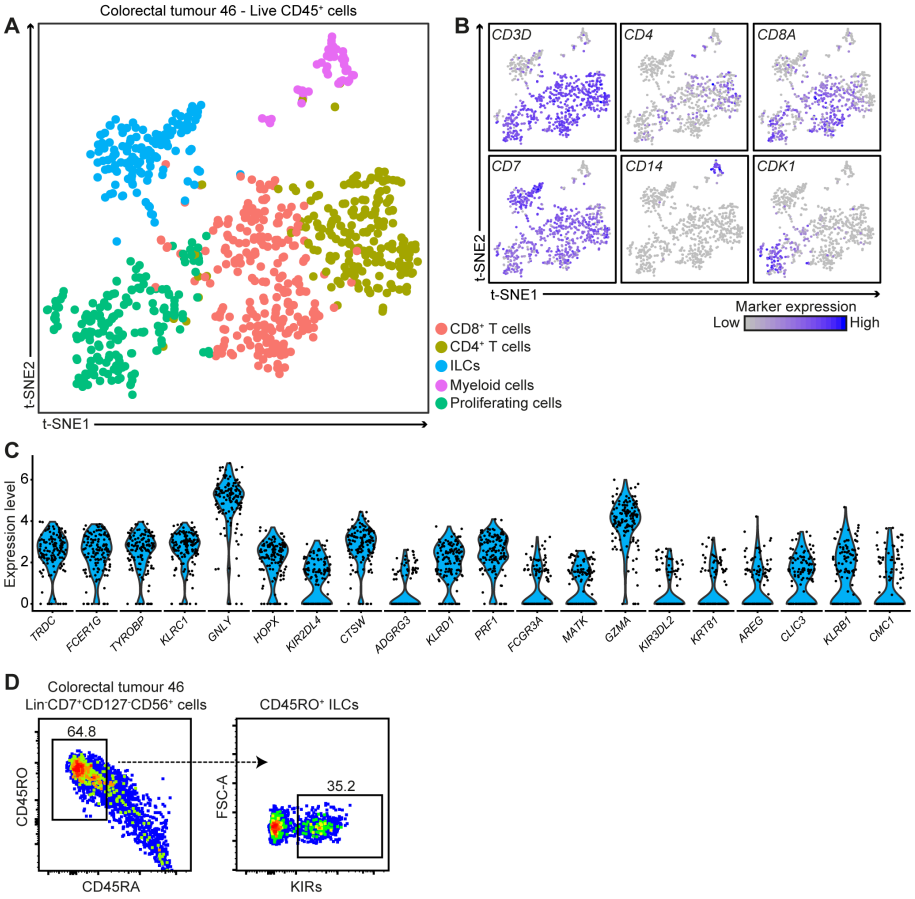


Supplementary Figure 6.5 Expression of FOXP3 by ICOS$^+$ regulatory T cells in colorectal tumours. (A) Representative plots of a healthy control PBMC sample and a MMR-proficient tumour sample analyzed by flow cytometry showing the expression of FOXP3 by regulatory T cells (CD25$^+$CD127^{low}). **(B)** FOXP3 expression in regulatory T cells (CD25$^+$CD127^{low}) from CRC tissues (N=4, of which 1 MMR-deficient and 3 MMR-proficient). Bars indicate median ± IQR. Each dot represents an individual sample. Data from two independent experiments with flow cytometry. **(C)** Representative plots of a healthy control PBMC

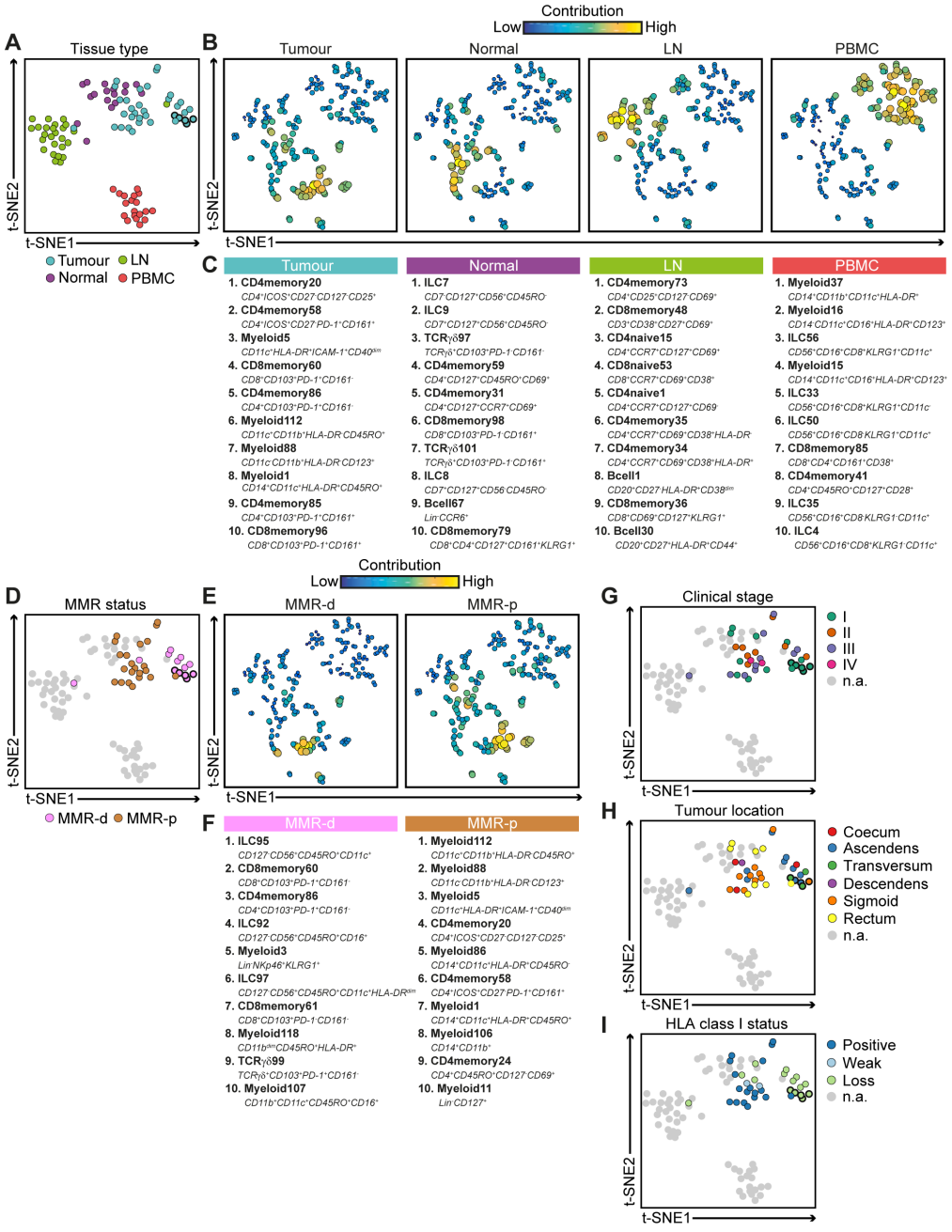
sample and a MMR-proficient tumour sample analyzed by flow cytometry showing the expression of FOXP3 by $ICOS^+$ regulatory T cells ($CD25^+CD127^{low}$). **(D)** FOXP3 expression in $ICOS^+$ regulatory T cells ($CD25^+CD127^{low}$) from CRC tissues (N=4, of which 1 MMR-deficient and 3 MMR-proficient). Bars indicate median ± IQR. Each dot represents an individual sample. Data from two independent experiments with flow cytometry. MMR-d; mismatch repair-deficient, MMR-p; mismatch repair-proficient.



Supplementary Figure 6.6 Co-expression of *ICOS*, *TNFRSF4 (OX40R),* and *TNFRSF18 (GITR)* on $CD4^+$ T cells in colorectal cancers. t-SNE embedding showing 1,079 cells from CRC tissues (N=7) analyzed by single-cell RNA-sequencing. Colors represent the log-transformed expression levels of indicated markers. Each dot represents a single cell. t-SNE; t-distributed stochastic neighbour embedding.

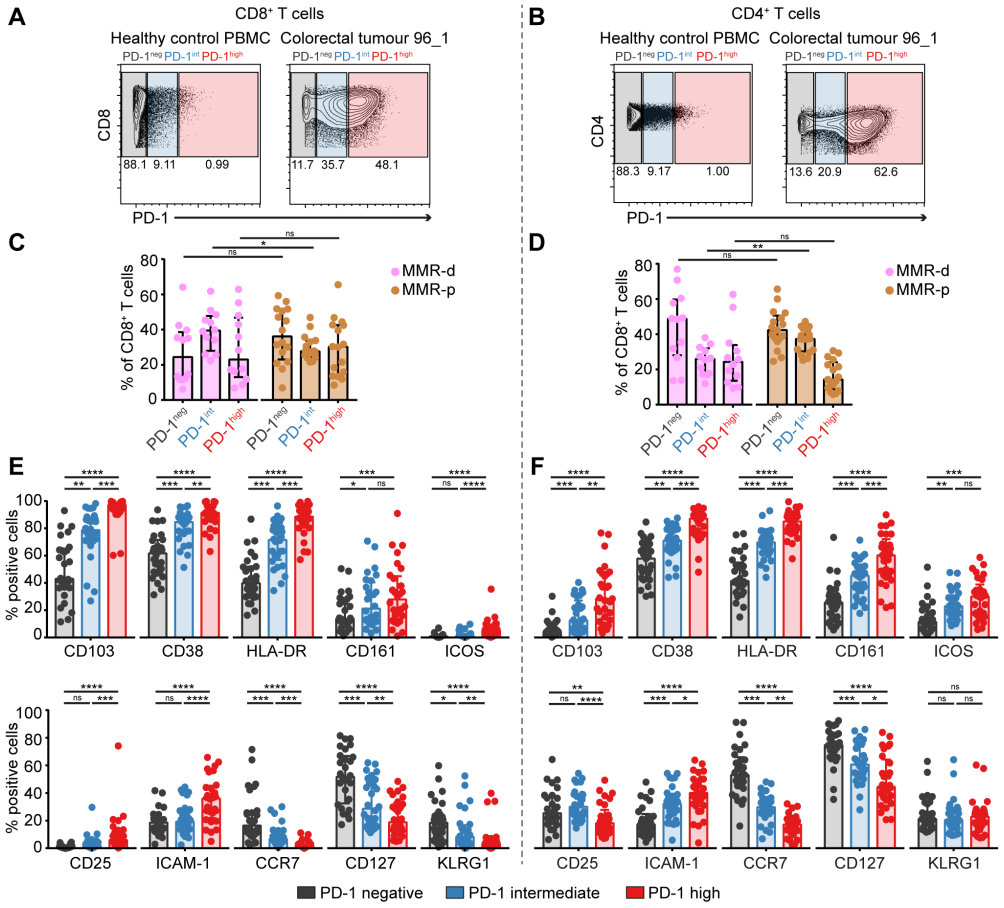


Supplementary Figure 6.7 Expression of cytotoxic molecules and KIRs by tumour-resident ILCs. **(A and B)** t-SNE embedding of single-cell RNA-sequencing data showing 795 cells from one MMR-deficient tumour that was selected for its high numbers of Lin$^-$CD7$^+$CD127$^-$CD56$^+$CD45RO$^+$ ILCs (70% of the ILC cluster) based on mass cytometry data. Colors represent the different clusters **(A)** and the log-transformed expression levels of indicated markers **(B)**. Each dot represents a single cell. **(C)** Violin plot showing log-transformed expression levels of the top 20 differentially expressed genes within ILCs (N=137) as identified in (A). Each dot represents a single cell. **(D)** Flow cytometry plots showing the cell surface expression of KIRs in Lin$^-$CD7$^+$CD127$^-$CD56$^+$CD45RO$^+$ ILCs from the same tumour as in (A-C). ILC; innate lymphoid cell, KIR; killer-cell immunoglobulin-like receptor, MMR; mismatch repair, t-SNE; t-distributed stochastic neighbour embedding.



Supplementary Figure 6.8 Integrated analysis of the immune composition in different tissue types of colorectal cancer patients. (A) Collective t-SNE analysis showing the clustering of 97 samples based on cell percentage data (of $CD45^+$ cells) of 218 immune cell clusters. Every dot represents a sample colored by tissue type. Five primary tumours at different locations from the same patient are highlighted. One lymph node sample clustered within the tumour samples, and was found to be infiltrated by tumour cells upon histological examination. One tumour sample clustered within the lymph node samples, and was found to contain large populations of naive $CD4^+$ T cells and B cells, which are enriched in lymph nodes. Histological examination of the tumour confirmed the presence of lymphoid aggregates with germinal

centers, a Crohn-like lymphoid reaction that can be a feature of MMR-deficient tumours. **(B)** Collective t-SNE analysis showing the clustering of 218 immune cell clusters based on cell percentage data (of $CD45^+$ cells) of 97 samples. Every dot represents an immune cell cluster. Dot color and size indicate the contribution of the immune cell cluster to the respective t-SNE sample signatures as shown in (A). **(C)** Top ten ranked immune cell clusters contributing to the t-SNE sample signatures as shown in (A). Unique cluster IDs and a short description of their phenotype are displayed. **(D)** Collective t-SNE analysis of (A) colored by MMR status of the tumour samples. **(E)** Collective t-SNE analysis of (B) showing the contribution of the immune cell clusters to the respective t-SNE sample signatures as shown in (D). **(F)** Top ten ranked immune cell clusters contributing to the t-SNE sample signatures as shown in (D). Unique cluster IDs and a short description of their phenotype are displayed. **(G-I)** Collective t-SNE analysis of (A) colored by clinical stage **(G)**, tumour location **(H)**, and HLA class I status **(I)**. HLA; human leukocyte antigen, ILC; innate lymphoid cell, LN; lymph node, MMR-d; mismatch repair-deficient, MMR-p; mismatch repair-proficient, PBMC; peripheral blood mononuclear cell, t-SNE; t-distributed stochastic neighbour embedding.



Supplementary Figure 6.9 PD-1 expression level of CD8⁺ and CD4⁺ T cells correlates with distinct states of activation and differentiation in colorectal tumours. (A and B) Representative plots showing the gating strategy for PD-1 negative, intermediate and high CD8⁺ T cells **(A)** and CD4⁺ T cells **(B)** in healthy control PBMC and colorectal tumour tissues (see Supplementary Methods). **(C)** Frequencies of PD-1 negative, intermediate and high CD8⁺ T cells in MMR-deficient (N=12) and MMR-proficient (N=16) tumours. **(D)** Frequencies of PD-1 negative, intermediate and high CD4⁺ T cells in MMR-deficient (N=12) and MMR-proficient (N=17) tumours. In **C** and **D** bars indicate median ± IQR. Data from 22 independent experiments with mass cytometry. NS, not significant, **P*<0.05, ***P*<0.01 by Mann-Whitney U-test. **(E)** Frequencies of selected immune cell markers expressed by PD-1 negative, intermediate and high CD8⁺ T cells in CRCs (N=28). **(F)** Frequencies of selected immune cell markers expressed by PD-1 negative, intermediate and high CD4⁺ T cells in CRCs (N=29). In **E** and **F** bars indicate median ± IQR. Each dot represents an individual sample. Data from 22 independent experiments with mass cytometry. NS, not significant, **P*<0.05, ***P*<0.01, ****P*<0.001, *****P*<0.0001 by Friedman test with Dunn's test for multiple comparisons. MMR-d; mismatch repair-deficient, MMR-p; mismatch repair-proficient**,** PBMC; peripheral blood mononuclear cell.

CHAPTER 7

CyTOFmerge: Integrating Mass Cytometry Data Across Multiple Panels

Tamim Abdelaal

Thomas Höllt

Vincent van Unen

Boudewijn P.F. Lelieveldt

Frits Koning

Marcel J.T. Reinders

Ahmed Mahfouz

This chapter is published in: *Bioinformatics* (2019) 35(20): 4063–4071, doi: 10.1093/bioinformatics/btz180.
Supplementary material is available online at: https://academic.oup.com/bioinformatics/article/35/20/4063/5381543#supplementary-data

High-dimensional mass cytometry (CyTOF) allows the simultaneous measurement of multiple cellular markers at single-cell level, providing a comprehensive view of cell compositions. However, the power of CyTOF to explore the full heterogeneity of a biological sample at the single-cell level is currently limited by the number of markers measured simultaneously on a single panel.

To extend the number of markers per cell, we propose an *in silico* method to integrate CyTOF datasets measured using multiple panels that share a set of markers. Additionally, we present an approach to select the most informative markers from an existing CyTOF dataset to be used as a shared marker set between panels. We demonstrate the feasibility of our methods by evaluating the quality of clustering and neighborhood preservation of the integrated dataset, on two public CyTOF datasets. We illustrate that by computationally extending the number of markers we can further untangle the heterogeneity of mass cytometry data, including rare cell-population detection.

7.1 INTRODUCTION

High-dimensional mass cytometry by time-of-flight (CyTOF)[1] allows the simultaneous measurement of over 40 protein cellular markers[2]. Several studies have illustrated the value of using such a large number of markers to provide an system-wide view of cellular phenotypes at the single-cell level[3–10].

Despite the three-fold extension in the set of markers profiled with CyTOF compared to flow cytometry (FC), technical challenges in designing CyTOF panels limit the number of markers profiled per panel currently to about 40 markers[11]. In many cases, the number of proteins required to describe the heterogeneity of cells far exceeds the number of markers that can be measured using a single CyTOF panel[10,12]. To overcome the limitation in the number of markers that can be measured simultaneously, a sample can be split into multiple tubes which are subsequently measured using different CyTOF marker panels[13–15]. Including a shared marker set between all panels allows the combination of measurements from all panels to produce an extended marker vector for each cell. However, there are currently no computational methods available to integrate measurements from multiple CyTOF panels.

An implicit combination approach, proposed by[12], allowed the visualization of 49 markers, measured using two CyTOF panels sharing 13 markers. After clustering cells from one panel based on the set of shared markers, they overlaid the unique markers of the second panel over the obtained clusters according to the similarity between cells based on the shared markers set. This approach, however, does not explicitly merge the measurements from both panels since the clustering step is performed only on cells from one panel using the shared markers. Therefore, this approach is prone to misidentify small subpopulations of cells (as we will show later in section 7.3.4).

In the field of Flow Cytometry (FC), two approaches have been proposed to integrate measurements from multiple FC datasets. A nearest neighbor algorithm was used to integrate measurements from multiple FC panels assuming that each cell is almost identical to its nearest neighbor cell, measured with a different panel, based on the overlapping markers, which we denote as the **first-nearest-neighbor** imputation[13,16,17]. However, the first-nearest-neighbor approach is noise-sensitive and can produce false combinations between cells from different panels resulting in artificial clusters[15]. Lee *et al.*, 2011 proposed to overcome this limitation by incorporating a clustering step based on the shared markers before merging the FC measured panels, followed by enforcing the imputation of the missing

markers from the same cluster, which we refer to as **cluster-based imputation**. However, the larger number of unique markers per panel in the case of CyTOF, compared to FC, can result in a large number of undiscovered clusters if cells are clustered only using the set of shared markers (as we will show later in section 7.3.2). An alternative approach is to divide the space of shared markers in each panel by binning biaxial scatter plots of marker pairs, each having a pre-set number of cells. Bins are then matched across the measured panels, and the missing markers are imputed per bin[15]. Although feasible for FC data, applying this method to CyTOF data, which has many more possible shared markers and many more cells, is computationally prohibitive. Moreover, the imputation strongly depends on the binning and matching step in a complex high-dimensional space.

We propose a method, CyTOFmerge, that does not depend on a priori clustering or partitioning and extends measurements per cell. Our CyTOF data merging approach is based on the k-nearest-neighbor algorithm which avoids the noise sensitivity problem by relying on a relatively large number of neighbors. In addition, we propose a method to select the most informative markers from one CyTOF panel, in order to be used as shared markers with other panels. This is particularly important given that the imputation strongly depends on the set of shared markers. By merging measurements from multiple CyTOF panels, we increase the number of markers per cell allowing for a deeper interrogation of cellular composition.

7.2 METHODS

7.2.1 APPROACH

Given that the maximum number of markers on a single CyTOF panel is N, the goal of our study is to integrate measurements from two CyTOF panels, panels A and B, given that both panels share at least $m < N$ markers. The remaining slots (N-m) on each panel can be used to measure markers that are unique to each panel. Both panels A and B measure parts of the same sample. Relying on the similarities between cells in both panels based on the shared marker set m, we can impute markers that were not measured on panel A using the measurements from panel B, and vice versa. The resulting merged dataset extends the number of markers per cell to $2N$-m, on which clustering and cell populations identification can be applied (Figure 7.1). We defined a *cell population* as group of cells having similar protein marker expression, these cells can represent either cells with the same type and/or state, according to which protein markers are used[18].

A major challenge in this approach is to determine the shared markers (m), i.e. which markers can preserve the heterogeneity of cell populations. To address this problem, we propose a data-driven approach (Supplementary Figure 7.1). Briefly, for each value of m, we use a dimensionality reduction technique to select the best set of markers preserving the high dimensional structure of the data. By simulating the scenario shown in Figure 7.1, the quality of an imputation is evaluated using several quantitative scores capturing clustering and neighborhood preservation, from which the minimum number of shared markers can be deduced. Full details of the selection process are described in section 7.2.6.

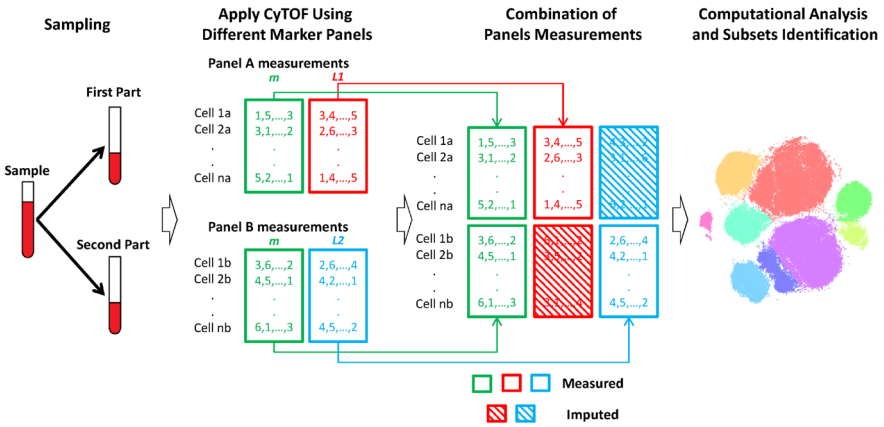


Figure 7.1 CyTOFmerge pipeline: Split the sample, stain each partial sample with a different marker panel and apply CyTOF to obtain the panels' measurements. Both panels A and B share a set of markers *m* (green). *L1* (red) are unique markers of panel A, and *L2* (blue) are unique markers of panel B. Both panel measurements are combined to obtain an extended markers measurements per cell, which is input to downstream computational analysis as, for example, clustering in a t-SNE mapped domain shown here.

7.2.2 CyTOF Datasets

In this study, we applied our methods to the publicly available HMIS and Vortex data sets. The HMIS data set profiled the human mucosal immune system by measuring Peripheral Blood Mononuclear Cells (PBMCs) and intestine tissue samples from the duodenum, rectum and fistula[8]. Using a CyTOF panel with N = 28 surface protein markers, a total of ~5.2 million cells positively expressing CD45 (immune cell marker) were analyzed (3.6 million PBMCs and 1.6 million intestine tissue cells), which they down sampled to ~1.1 million cells, randomly distributed over all PBMC and tissue cells. The marker panel included lineage markers used to differentiate between major types of immune cells, and non-lineage markers used to distinguish between different subgroups (states) of cells within each lineage. Cells were globally clustered into six main lineages: B cells (~ 93,000), CD4+ T cells (~ 230,000), CD8+ T cells (~ 460,000), CD3-CD7+ Innate lymphoid cells (ILCs) (~ 95,000), Myeloid cells (~ 117,000) and TCRγδ cells (~ 88,000). Each lineage was subsequently clustered independently, resulting in 119 subgroups across all six lineages, including small clusters representing rare cell populations.

The Vortex dataset is a publicly available mass cytometry data for 10 replicates of mice bone marrow cells[19]. A total of ~840,000 cells were measured using a CyTOF panel of N=39 markers. Three cytometry experts provided a consensus clustering of 24 clusters for only ~510,000 cells. Prior to any processing, measured marker expressions were transformed using hyperbolic arcsin with a cofactor of 5 for both datasets.

7.2.3 Simulating Two Overlapping Panels

We simulated the scenario of having two overlapping panels by splitting the original dataset (D^o) into two datasets, D_A and D_B, each measured using a different (simulated) CyTOF panel (Supplementary Figure 7.1). Both panels share m markers, and the remaining N-m markers from the original panel were randomly divided between the two simulated panels. The first simulated panel (A) contains m+$L1$ markers, whereas the second panel (B) contains m+$L2$ markers, where $L1$+$L2$=N-m. Each of the two panels measures half the number of cells in the

original dataset (randomly chosen without replacement), i.e. the panels measure non-overlapping cells from the original dataset.

7.2.4 Data Imputation

Data in both simulated CyTOF panels is imputed using the *k*-nearest neighbor algorithm. For each cell measured by panel A, we find the *k*-most similar cells measured by panel B using the *m* shared markers. Then, for each cell measured by panel A, the values of the missing markers (*L2*) are imputed by taking the median values of those markers from the *k*-most similar cells measured by panel B, resulting in imputed dataset D_A^i. The same procedure is used to impute the values of the missing markers *L1* from panel A to cells measured with panel B, resulting in imputed dataset D_B^i. The original dataset is reconstructed (D^i) by concatenating the two imputed datasets (D_A^i and D_B^i), and thus has the same number of cells and the same number of markers *N* as the original dataset, albeit partly imputed (Figure 7.1 and Supplementary Figure 7.1).

7.2.5 Selection of *M* Shared Markers

Given a dataset with a panel of *N* markers, we follow three steps to choose the *m* shared markers that can be used to design follow up panels for a deeper interrogation of cells (Supplementary Figure 7.1):

Removing correlated makers. Pearson correlation over all cells in the original dataset between each pair of markers is calculated. If the absolute value of the correlation of two markers is larger than a specified cutoff (here we use 0.7 and 0.8 as cutoffs, for the HMIS and Vortex datasets, respectively), we remove the marker which has the lower variance across all cells.

***Dimensionality reduction*.** To reduce the number of markers we exploited three different dimension reduction techniques: (i) principal component analysis (PCA); (ii) Auto Encoder (AE) and (iii) Hierarchical Stochastic Neighboring Embedding (HSNE).

Using PCA[20], the importance of a marker is based on its contribution (i.e. loading factor) to the first *m* principal components, as follows:

$$i_p = \sum_{q=1}^{m} \beta_{pq}^2 * \lambda_q \tag{7.1}$$

where i_p is the importance of marker *p*, β_{pq} is the loading of marker *p* to the *q-th* Principle Component (PC), λ_q is the variance explained by the *q-th* PC. All markers are sorted on their importance and the *m* most important markers are chosen.

An auto encoder neural network[21] with one hidden layer containing *m* nodes is trained for a maximum of 50 iterations (using the Matlab toolbox for Dimensionality Reduction, drtoolbox: https://lvdmaaten.github.io/drtoolbox/) until the output of the trained auto encoder is similar (mean squared error < 0.75 for all values of *m*) to the original input data. We then calculate the variance of all auto encoder output markers, sort them and select the *m* markers with the highest variance.

Using Hierarchical Stochastic Neighboring Embedding (HSNE)[22,23], we project the cells using five hierarchical layers. We represent the dataset using only the landmark cells in the top

layer. On these landmark cells we apply the PCA-based reduction scheme to select the *m* markers.

Selecting m out of the original N markers. Using one of the dimension reduction schemes, we select the top-*m* markers to be used as shared markers. Based on the simulated datasets, we impute the missing markers in each dataset, which we compare to the original dataset using three quantitative scores introduced in the following section. By evaluating those scores over varying values for *m*, we make a choice for the most suitable value of *m*.

7.2.6 COMPARING TWO DATASETS

To evaluate the quality of the imputed dataset (D^i) compared to the original dataset (D^o), we use three different scores: (i) how well the clustering is preserved *(cluster score)*; (ii) how close the same cells in the different data sets are to each other *(distance score)* and (iii) how well the neighborhood of each cell is preserved *(nearest neighbor score)*. These scores are defined as follows:

***Cluster score*.** We used the adjusted Rand-index to express the correspondence between two clustering. Briefly, it calculates the fraction of pairs of cells that end up in the same (or different) cluster in both clusterings, corrected for the random chance to end up in the same cluster (which is different for differently size clusters). The final value is between 0 and 1. As clustering more than a million cells is too time consuming, we used an approximate cluster score for experiments where we varied either the number of shared markers (*m*) or neighbors used to impute (*k*). For these experiments, we did not cluster the imputed data D^i but determined the cluster label of the imputed cell by a majority vote of the *k* most-similar cells in the original data set D^o. The *approximate cluster score* is then the fraction of cells where the estimated cluster label was the same as the cluster label of the original cell:

$$\frac{Approximate}{Cluster\ Score} = \frac{number\ \ of\ cells\ having\ matched\ cluster\ labels}{total\ number\ \ of\ cells} \tag{7.2}$$

Distance score. To evaluate how similar the measurements of cells across two datasets are, we calculate the Euclidean distance, in the full marker space, between the measurements of a cell c_n^i, the *n*-th cell in the imputed dataset D^i, and the corresponding cell c_n^o, the same (*n*-th) cell in the original dataset D^o. This is done for all cells, and from that the median distance *(md)* is taken. To make the score independent of the scale of the original data set D^o, we compare this median distance *(md)* to the average distance *(ad)* between all pairs of cells within the original dataset D^o, as follows:

$$Distance\ Score = \frac{(ad\ -\ md)}{ad} \tag{7.3}$$

***Nearest Neighbor score*.** To evaluate the preservation of the neighborhood of cells across datasets, we measure, for each cell c_n^o, the Euclidean distance in the full marker space to the nearest neighboring cell (d_n) in the original dataset D^o, and the distance between both representations of that cell, c_n^o and c_n^i, in the original D^o and imputed D^i datasets (d_p). The local neighborhood is preserved when the imputed version of the cell c_n^i is closer to c_n^o than its nearest neighbor in the original dataset D^o, i.e. $d_p < d_n$. The nearest neighbor score is then the fraction of cells for which this holds.

$$NN\ Score = \frac{\text{number of cells where } (d_p < d_n)}{\text{total number of cells}} \quad (7.4)$$

We used the base 2 logarithm of the Jensen-Shannon divergence (JSD) to quantify the similarity between the distributions of a marker in the original and imputed dataset, resulting in values between zero (identical distributions) to one (totally disjoint distributions). The JSD between two distributions $P(x)$ and $Q(x)$ is:

$$JSD = \frac{1}{2}\sum_x P(x)\ log_2\left(\frac{P(x)}{M(x)}\right) + \frac{1}{2}\sum_x Q(x)\ log_2\left(\frac{Q(x)}{M(x)}\right) \quad (7.5)$$

$$M(x) = 0.5 * (P(x) + Q(x)) \quad (7.6)$$

7.2.7 FINDING CLUSTERS

We clustered both datasets, HMIS and Vortex, with Phenograph, a neighborhood graph-based clustering tool designed for automated analysis of mass cytometry data[24]. Phenograph is applied to the original and imputed datasets, using the R implementation with default settings (number of neighbors = 30).

More fine-grained cluster annotations for the HMIS datasets are acquired using Cytosplore (www.cytosplore.org), a tool specifically designed for the analysis of mass cytometry data[23,25]. Briefly, cells are embedded into a two-dimensional map using t-Distributed Stochastic Neighbor Embedding (t-SNE)[26,27], and subsequently clustered using a density-based Gaussian Mean Shift (GMS) algorithm[28] using a relatively small density kernel (σ = 20-23), resulting in over-clustering of the data. Clusters are then manually merged when they have highly similar marker expression profiles (median value of each marker per cluster).

7.3 RESULTS

7.3.1 SELECTING THE SET OF SHARED MARKERS

To determine the shared markers that can be used to combine two CyTOF datasets, we simulated the scenario of having two overlapping panels with different sets of shared markers *m,* on which we applied our data imputation approach with different number of neighbors *k* (Supplementary Figure 7.1). We investigated how the imputation of the two panels is influenced by: (i) the dimension reduction technique used to select the shared markers, (ii) the data (lineages) used to select the markers, (iii) the number of shared markers (*m*), and (iv) the number of nearest neighbors used during imputation (*k*).

In the HMIS dataset, the method used to select the shared markers has limited influence on the results. Figure 7.2 shows which markers are selected by the different marker selection schemes (PCA, AE and HSNE) when changing the number of selected shared markers (*m*) from 4 to 25 and applied on the 5.2 million cells. In the pre-processing step, CD8b and CD11b were removed from the selection as they are highly correlated with CD8a and CD11c (correlation of 0.843 and 0.705, respectively), leaving 26 markers to choose from. There are small differences in the selection profiles between the three methods, with a maximum of two mismatches. For $14 < m < 17$, the same set of shared markers is selected by all three methods. In terms of computation time, PCA outperforms the Auto Encoder and the HSNE (100x and 480x, faster on the same machine, respectively).

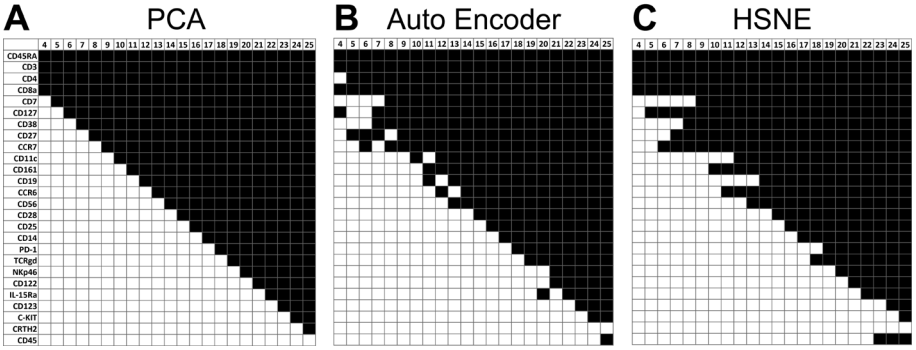


Figure 7.2 Shared markers for the HMIS dataset. The selected markers that can best represent the dataset using **(A)** PCA, **(B)** Auto Encoder and **(C)** HSNE. (Marker ordering is based on the PCA selection profile, black is selected, white is not selected).

We checked whether the marker selection procedure is influenced by the type of cells. Therefore, we applied the PCA-based marker selection on PBMCs and tissue cells independently. Supplementary Figure 7.2 shows that there is little difference in the selected set of markers when using the PBMC, tissue or PBMC+tissue samples.

Next, we assessed the quality of the subsequent imputed dataset for each lineage individually, as well as all six lineages together, for m = 4 to 25 and k = 50. For all three evaluation scores, the performances improve when the number of shared markers increases (Supplementary Figure 7.3A-C). All performance scores seem to saturate at m = 16 (Supplementary Figure 7.4A-F), i.e. they exceed 80% of the maximal score. Table 7.1 shows the values of the three quality measures at m = 16, for each individual lineage and the six lineages together.

A common measure to assess the quality of imputation is to investigate the correlation between the original and imputed values. However, this approach turned out not to be appropriate for our data since many markers are being expressed only in a specific population of cells. As a result, the correlation is relatively high for markers that are high expressed over multiple cell populations (Supplementary Figure 7.5 and 7.6), but the correlation is low for cell-population specific markers (such as, for example, the CD123 marker which is high expressed only in the CD4+ T cells lineage). These cell-population specific markers are imputed correctly (low values for most cells and higher values for the cell-population specific cells), but the noise on the abundant low values dominates, causing a low correlation. Consequently, we decided not to use the correlation as a quantitative score to evaluate how well an imputed dataset resembles an original dataset.

Table 7.1 Evaluation scores for the 16 selected shared markers for the 1.1 million cells HMIS dataset.

	Approximate Cluster Score	*Distance Score*	*Nearest Neighbor Score*
CD4+ T Cells	92.3 %	84.3 %	94.5 %
CD8+ T Cells	91.9 %	83.9 %	93.1 %
B Cells	91.8 %	82.0 %	92.8 %
CD3-CD7+ Cells	89.3 %	83.4 %	92.6 %
TCRγδ Cells	86.2 %	84.1 %	94.7 %
Myeloid Cells	86.2 %	80.4 %	82.5 %
All Cells	89.4 %	87.4 %	91.9 %

We further investigated the distribution of the non-shared (imputed) marker by comparing the distributions of the original values with those of the imputed values for each non-shared marker per cell population, and quantify the similarity using the JSD (Methods). Across all the 12 non-shared markers, we obtained low JSD values (<0.2) showing a high similarity between the original and imputed values (Supplementary Figure 7.7A). The imputation process does exclude the outlier values, as we use the median value from the 50 most similar cells, which results for some markers, in 'compressed' distributions as compared to the original ones (Supplementary Figure 7.7B-C).

Next, we investigated the effect of the choice of the number of neighbors (k) used when applying the k-nearest neighbor imputation. Supplementary Figure 7.4A-F shows the *approximate cluster score* for k = {1, 10, 50, 100, 200, 250, 300, 500, 1000}, with k = 50 clearly showing the highest performance across all lineages, even over different numbers of shared markers.

We observed similar results when applying all these analyses to the Vortex dataset: (i) small differences between PCA, AE and HSNE when m is ranging from 4 to 38 (Supplementary Figure 7.8), (ii) improving and saturating performance scores with increasing number of shared markers (Supplementary Figure 7.3D), and (iii) highest performance when k=50 is used during imputation (Supplementary Figure 7.4G). The saturation for the number of shared markers occurs at m = 11, with the *approximate cluster score*, *distance score* and *nearest neighbor score* being 95.3%, 84.0% and 82.1%, respectively.

7.3.2 CyTOFmerge Reproduces Original Cell Populations and Outperforms FC Imputation Methods

To demonstrate the feasibility of our computational method to combine data measured from multiple CyTOF panels, we investigated the quality of the clustering of the imputed dataset. First, the original 1.1 million cells HMIS dataset was clustered on the full marker space using Phenograph, resulting in 52 clusters of cells divided into: 6 B cell populations, 8 CD4+ T cell populations, 15 CD8+ T cell populations, 6 CD3-CD7+ ILC populations, 7 Myeloid populations, 5 TCRγδ cell populations and 5 unknown populations donated as Others (Supplementary Figure 7.9). These 52 clusters are used as a baseline for comparison with the imputed datasets.

We applied the panel combination and imputation method using k = 50 and m = 16, thus imputing 12 markers (6 unique markers for panel A, and 6 unique markers for panel B). The imputed dataset was clustered on the full marker space using Phenograph, resulting (coincidentally) in 52 clusters with slight variation in the number of clusters per cell lineage (Supplementary Figure 7.10A). To evaluate the imputation, we matched the imputed clusters to the original clusters using the maximum pairwise Jaccard index. The cluster matching shows that all imputed clusters match to original clusters within the same lineage (Supplementary Figure 7.10B). Next, we calculated the adjusted Rand-index representing how similar both clusterings are (Table 7.2).

To compare with the first-nearest-neighbor approach proposed by[13], we applied the imputation method using k = 1, using the same set of 16 shared markers. Phenograph clustering of that imputed dataset on the full marker space resulted into 53 clusters (Supplementary Figure 7.11) with a lower performance compared to CyTOFmerge using k = 50 (Table 7.2).

Table 7.2 Comparison between CyTOFmerge and FC merging methods on the 1.1 million cells HMIS dataset.

	Adjusted Rand-index	*Distance Score*	*Nearest Neighbor Score*
CyTOFmerge			
HMIS, m = 16, k = 50	0.81	87.4 %	91.9 %
Vortex, m = 11, k = 50	0.90	84.0 %	82.1 %
First-nearest-neighbor			
HMIS, m = 16, k = 1	0.77	83.5 %	75.6 %
Vortex, m = 11, k = 1	0.93	77.9 %	51.6 %
Shared markers clusters			
HMIS, m = 16	0.68	n.a	n.a
Vortex, m = 11	0.79	n.a	n.a
Cluster-based imputation			
HMIS, m = 16, k = 50	0.80	87.4 %	91.8 %
Vortex, m = 11, k = 50	0.84	84.0 %	82.1 %

n.a = not applicable

Next, we compared the performance of CyTOFmerge to that of the cluster-based imputation method proposed by[14]. In this approach, clusters are first determined using the shared markers followed by imputation of the unique markers in each panel *within* the same cluster. We clustered the cells using the 16 shared markers for the entire dataset using Phenograph and obtained 42 cell clusters, 10 clusters less than the original dataset clustering (Supplementary Figure 7.12). When comparing with the original clustering (Table 7.2), we observed a relatively large drop in the adjusted Rand-index. Hence, clustering based on the shared markers only could not identify a large part of the original clustering using all markers. However, when we performed the combination of the two panels using the cluster-based imputation, we obtained comparable performance with CyTOFmerge (Supplementary Figure 7.13, Table 7.2).

We also tested CyTOFmerge on the Vortex dataset, using m = 11 shared markers and k = 50, now imputing 28 markers (14 unique per panel). Phenograph clustering of the original dataset gave 31 clusters (Supplementary Figure 7.14), while clustering the imputed dataset resulted in 28 clusters (Supplementary Figure 7.15). The adjusted Rand-index was relatively high, i.e. 0.90 (Table 7.2). Next, we applied first-nearest-neighbor approach, and we clustered the resulting imputed dataset resulting in 29 clusters. The first-nearest-neighbor has slightly higher adjusted Rand-index compared to CyTOFmerge, however, we observed a large drop in the *distance* and the *nearest-neighbor scores* (Table 7.2). Moreover, confirming our previous observation, the clustering of the shared markers only produces 23 clusters, 8 clusters less than the original dataset clusters, with a relatively large drop in the adjusted Rand-index when compared to the original clustering. Finally, the cluster-based imputation method produces 29 clusters. Compared to CyTOFmerge, the cluster-based imputation method shows comparable *distance* and *nearest-neighbor scores,* but lower adjusted Rand-index (Table 7.2).

To obtain a baseline evaluation for the imputed data clustering performance, we permutated the non-shared markers across all cells, while keeping the shared markers values the same. Next, we clustered this permuted dataset in the full marker space using Phenograph and compared the clustering result with the original dataset clustering. The permuted dataset clustering had an adjusted Rand-index of 0.56 ± 0.02 and 0.50 ± 0.01 (across 10 different random permutation), for the HMIS and Vortex datasets, respectively. These results show

that random estimation of the non-shared markers decreases the clustering performance compared to clustering using the shared markers only, i.e. adding more dimensions does not improve the clustering performance. This also implies that CyTOFmerge adds real structure by providing good estimation for the non-shared markers, leading to an improved clustering.

7.3.3 Reproducible Cell Populations at a Deeper Annotation Level using CyTOFmerge

We proceeded by evaluating the quality of CyTOFmerge when using a fine-grained clustering to investigate whether rare (small) cell populations could be identified from the imputed data. As a baseline for comparison, we clustered the six immune lineages from the original 1.1 million cells HMIS dataset individually, on the full marker space using Cytosplore, resulting in 121 clusters in total, including: 17 CD4+T cell populations, 21 CD8+ T cell populations, 16 B cell populations, 34 TCRγδ cell populations, 24 CD3-CD7+ ILC populations and 9 Myeloid cell populations (Figure 7.3A, Supplementary Figure 7.16A). The imputed dataset (with m = 16) was similarly clustered using Cytosplore into the same number of populations (121) for the six immune lineages (Figure 7.3B, Supplementary Figure 7.16B).

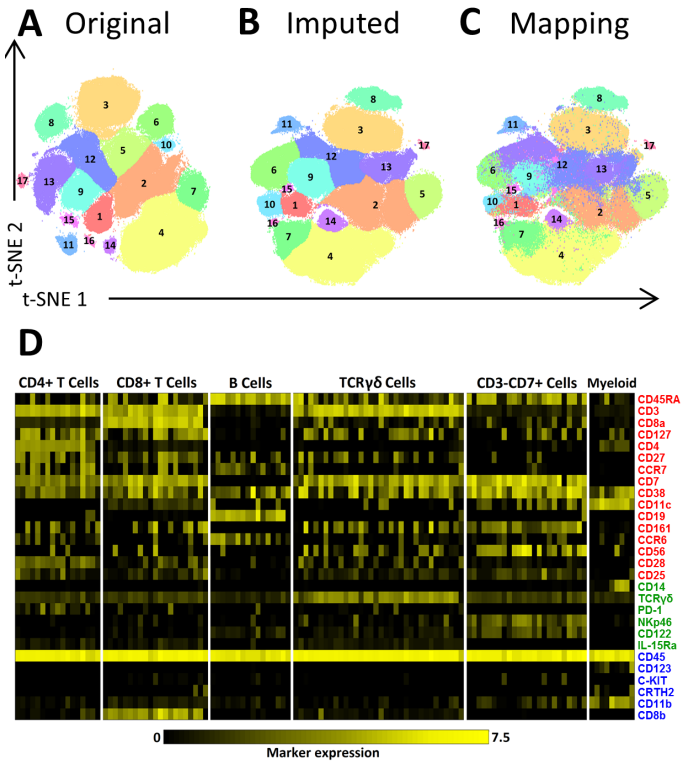


Figure 7.3 Clustering of the original and the imputed datasets. (A-C) t-SNE maps showing the different identified populations in the CD4+ T Cells lineage. **(A)** shows the populations of the original data. **(B)** The populations of the imputed data (for m=16, $L1$=6 and $L2$=6). **(C)** The mapping of the original clusters labels on the t-SNE map of the imputed data. **(D)** Heatmap of markers expression for the 121 characterized immune cells populations of the original dataset for m = 16. Black-to-yellow scale shows the median arcsinh-5 transformed values for the markers expression. Markers colors indicate whether a marker is shared between panels or unique to a single panel, during panels combination (red is shared, green is unique to panel A, blue is unique to panel B).

The clusters from the imputed dataset were correctly matched to the baseline clusters for all 121 cell populations across the six lineages, including large clusters as well as small rare clusters, such as: population 16 and 17 in the CD4+ T Cells (Figure 7.3A-B), population 21 in the CD8+T Cells, population 16 in the B Cells, populations 3 and 34 in the TCRγδ Cells, and populations 23 and 24 in the CD3-CD7+ Cells (Supplementary Figure 7.16A-B). The imputed expression profiles of the 121 populations are remarkably similar (average correlation of 0.998) to the expression profiles of the corresponding baseline clusters (Supplementary Figure 7.17A and Figure 7.3D, respectively). Also, the Jaccard index showed a clear diagonal between the original and the imputed clusters (Supplementary Figure 7.18).

To gain more insight into the distribution of the original cluster labels in the imputed space, we colored each cell in the imputed data according to baseline cluster they belonged to. Figure 7.3C and Supplementary Figure 7.16C show that the imputed measurements for each cell are indeed faithfully reconstructed, i.e. after mapping them they are distributed similarly as in the original data.

More quantitatively, the imputation had an overall adjusted Rand-index of 0.81 for all the 121 cell populations. Per individual lineage, the adjusted Rand-index varied between 0.77 and 0.83 for the different lineages (Table 7.3). Since we rely on GMS clustering in the t-SNE space, part of the error in clustering the imputed data is caused by the stochastic nature of the t-SNE algorithm (due to random initializations). The clustering reproducibility between two t-SNE mappings of the original data (Table 7.3, Supplementary Figure 7.19) varied between 0.82 and 0.96, with variance estimates (when repeating the procedure 10 times) in the order of 8e-5 (Table 7.3, for Myeloid and TCRγδ cells). Hence, the quality of the imputed clustering is close to the quality of repeated t-SNE mappings, with a difference of 0.06 in the adjusted Rand-index for all cells.

To further evaluate the effects of imputation on downstream analysis, we compared the population frequencies of the 121 cell populations, estimated using both the original and the imputed datasets. The result shows that population frequencies are accurately estimated from the imputed data as compared to the original data, with an overall correlation of 0.985 (Supplementary Figure 7.17B).

Table 7.3 Adjusted Rand-index of the imputed data at m = 16 and for repeated t-SNE mappings of the original data.

	Imputed data	t-SNE rerun
CD4+ T Cells	0.78	0.86
CD8+ T Cells	0.79	0.84
B Cells	0.83	0.85
CD3-CD7+ Cells	0.78	0.82
TCRγδ Cells	0.77 ± 8e-5	0.89 ± 1e-4
Myeloid Cells	0.82 ± 7e-5	0.96 ± 6e-5
All Cells	0.81	0.87

7.3.4 IMPUTATION IMPROVES THE DIFFERENTIATION OF CELL POPULATIONS

We have shown that from the imputed data similar clusters of cells can be found as when using the original data. But, can we find clusters from the imputed data that we cannot find in the two separate panels? Hereto, we overlaid the original cluster labels of the HMIS TCRγδ lineage populations onto t-SNE maps constructed using: (i) only the 22 measured markers of a panel (16 shared + 6 unique markers), (ii) the original 28 measured markers, and (iii) the imputed dataset (16 shared + 6 unique + 6 imputed). This was done for both panels A and B separately (Figure 7.4 and Figure 7.5, respectively).

For panel A, populations 6 and 8 are merged in one cluster when we map the data using only the 22 panel markers (Figure 7.4A), whereas the original and imputed data separate those two clusters (Figure 7.4B-C, respectively). To better understand this behavior, we overlaid the expression of the markers across the t-SNE map (Figure 7.4D). CD8b has higher expression (mean±std = 3.205±0.797) for cells in cluster 6 as compared to cluster 8 (0.584±0.663) and is missing in panel A, hence resulting in not being able to separate clusters 6 and 8. For the imputed data, the missing marker for panel A is imputed by its measurements on panel B, with which both clusters can indeed be separated (Figure 7.4C).

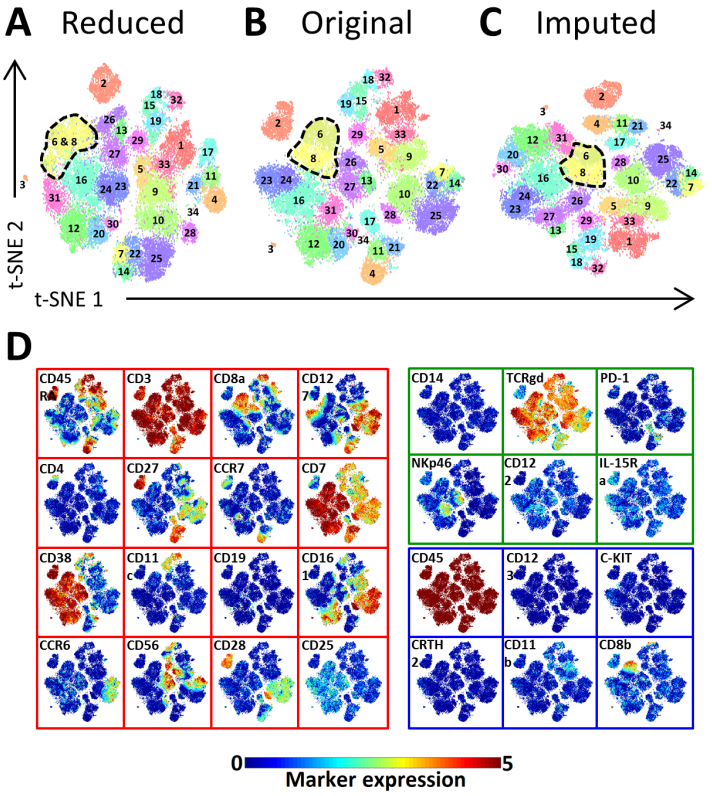


Figure 7.4 Marker panel extension impact on the identification of distinct populations in the TCRγδ immune lineage – Panel A. (A) The Reduced t-SNE map using only 22 markers. **(B)** The original t-SNE map using the original 28 markers. **(C)** The imputed t-SNE map using 28 markers of which 6 are imputed from Panel B). All three maps are colored with the original population labels. **(D)** Shared and missing markers expression profiles are shown on the original t-SNE map. The map border color indicate whether a marker is shared between panels or unique to a single panel (red is shared, green is unique to panel A, blue is unique to panel B and thus missing markers for panel A).The color bar shows the arcsinh-5 transformed values for the markers expression.

Likewise, for the data from panel B, cluster 12 and 31 are merged in one cluster (Figure 7.5A), because NKp46 is missing on panel B (Figure 7.5D) with cells having a higher expression in cluster 31 (2.728±0.712) compared to 12 (0.505±0.586). Also, clusters 7 and 14 are merged due to the lack of the TCRγδ marker (Figure 7.5D). For both situations, the clusters are separated when the data from panel B is imputed with data from panel A (Figure 7.5C).

Similar observations can be made for the other lineages (Supplementary Figure 7.20 – 7.24). For example, for both the CD8+ T (Supplementary Figure 7.20) and Myeloid (Supplementary Figure 7.21) lineages, the CRTH2 marker makes a difference between clusters based on one panel-only data compared to data from combined panels. For some lineages, the clustering based on individual panels does, however, closely match the clustering on the original data. Either the missing markers are not important (e.g. CD11b in panel A of the CD8+ T cells, Supplementary Figure 7.20), or they are important but highly correlated with one of the shared markers (e.g. CD14 in panel B of the Myeloid cells, Supplementary Figure 7.21, has a similar expression to CD38).

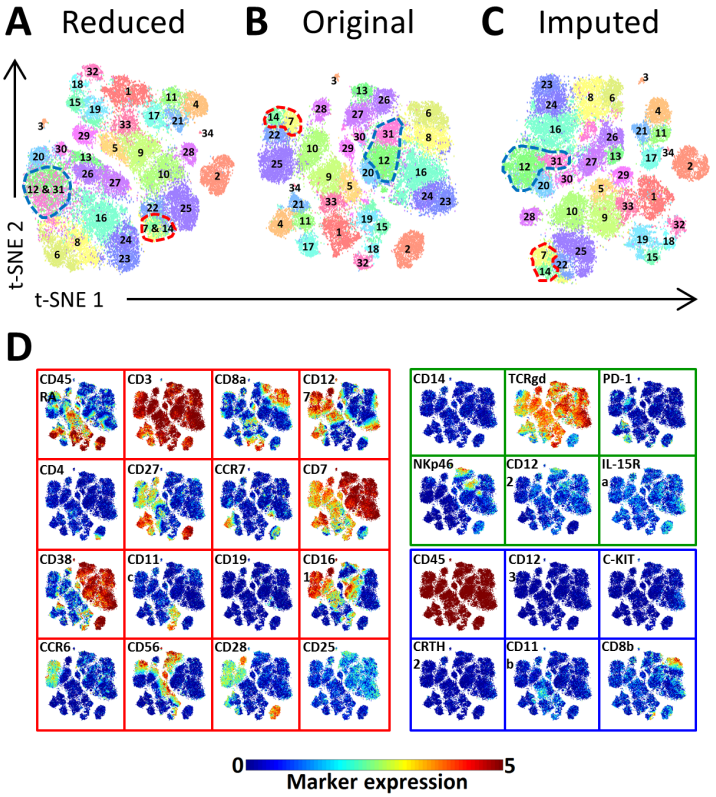


Figure 7.5 Marker panel extension impact on the identification of distinct populations in the TCRγδ immune lineage – Panel B. (A) The Reduced t-SNE map using only 22 markers. **(B)** The original t-SNE map using the original 28 markers values. **(C)** The imputed t-SNE map using 28 markers of which 6 are imputed from panel A. All three maps are colored with the original populations labels. **(D)** Shared and missing markers expression profiles are shown on the original t-SNE map. The map border color indicate whether a marker is shared between panels or unique to a single panel (red is shared, green is unique to panel A and thus missing markers for panel B, blue is unique to panel B).The color bar shows the arcsinh-5 transformed values for the markers expression.

To quantitatively assess the ability to differentiate between cell populations based on different sets of markers, we tested the ability of a two-class Linear Discriminant Analysis (LDA) classifier[29], to differentiate between populations 6 and 8 in the TCRγδ cells. We evaluated LDA's performance using only the 16 shared markers, all 28 markers from the TCRγδ imputed data, and all 28 markers from the TCRγδ original data. We obtained the highest performance using all markers from the original data, with an accuracy of 95.74 ± 0.70%. The lowest performance was obtained when using only the 16 shared markers (accuracy = 70.37 ± 1.07%). Using all markers from the imputed data resulted in an accuracy of 83.46 ± 1.13%, which is less than the original data, as expected, but showing a strong improvement over the shared markers. This confirms our previous conclusion that the imputation improves over the shared markers, despite the fact that the imputation relies on the shared markers. We obtained similar results for populations 12 and 31, and populations 7 and 14 (Supplementary Figure 7.25).

7.4 Discussion

We demonstrated the feasibility of combining data from different CyTOF panels with a set of shared markers in common. We showed that by imputing data, the heterogeneity of the data can be better captured than with the individual panels separately. Also, we presented a data-driven approach to select the set of shared markers that are most informative to be used to align panels.

The selected set of shared markers can capture the underlying structure of the data. For example, from the HMIS dataset we saw that for small values of m, the selected shared markers include CD3, CD4 and CD8a which separate the main CD4+ and CD8+ T cells immune lineages from the rest of the cell populations. As m increases, the selection algorithm starts to include markers that differentiate the different populations within a single lineage. Our selection approach relies on the variation in expression across cells. As a result, CD45, an essential marker which is positively expressed across all immune cells, is never selected due to its low variance.

To assess the quality of imputation, we relied on three scores that capture the cluster and neighborhood concordance between the imputed and original data. For the HMIS dataset, we observed prominent discordance when a low number of shared markers is used ($m < 12$), mainly due to exclusion of key lineage specific markers within the set of shared markers resulting in imputation failures. The number of shared markers to properly align panels does depend heavily on the complexity and heterogeneity of the data. For the HMIS dataset, studying PBMCs and tissue samples from patients with three different inflammatory bowel diseases as well as controls, 16 shared markers were needed. Whereas for the Vortex dataset, that replicated mouse bone marrow samples, 11 markers were sufficient. On the other hand, we saw that for both datasets we can capture and reconstruct all cell clusters, despite their number and sizes, suggesting that the imputation is not biased towards the clustering. Although the performances do differ for different settings of the number of shared markers (m) and number of neighbors used during imputation (k), they are not sensitive to the exact setting, illustrating the robustness of CyTOFmerge.

Note that during the shared maker selection procedure we represented highly correlated markers by only one representative marker. We made this choice because highly correlated markers will get the same importance by the PCA selection scheme, and thus might be selected together. Selecting a highly correlated marker as an additional shared marker will, however, not add any information to the shared makers, while, at the same time, occupying

a marker slot in the panel. To reduce this redundancy and free as many slots as possible on the panel we made the choice to represent highly correlate makers with only one marker. Clearly, the choice for the threshold plays an important role as when the correlation is lower the markers will also add more distinct information.

We have shown that by imputing more markers, it is possible to better differentiate between cell populations, but on the other hand, the imputation of markers does affect the quality of the downstream analysis when compared to non-imputed data. We saw that clustering of the imputed data is not perfectly similar to the original data (adjusted Rand-index < 1). Indeed, this is affected by the homogeneity of the dataset, as we saw higher performance for the Vortex datasets compared to HMIS (Vortex being more homogenous). Generally, the number of shared markers will affect the downstream analysis, i.e. increasing the number of shared markers will increase the quality of the imputation, and the downstream analysis will more faithfully resemble analyses done on non-imputed data. But that will also restrict the number of unique marker slots available on each panel. Using less shared markers will increase the number of unique markers, which in turn will increase the capacity to capture more heterogeneity, but at the expense of imputation quality. This trade-off is being influenced by the local structure (homogeneity) in the data, which is, unfortunately, hard (or even impossible) to predict beforehand, in general.

Compared to FC methods, CyTOFmerge outperformed the first-nearest-neighbor method, and achieved comparable performance with the cluster-based imputation. The later shows that the pre-clustering step of the shared markers is unnecessary, as the imputation through the entire data using CyTOFmerge produces similar results. Further, we demonstrated that by imputing more markers, we obtained better differentiation between different cell populations. However, the imputation depends on how similar cells are in the shared markers space, indicating that the variation between populations that can only be differentiated based on imputed (non-shared) markers is to some extent retained in the shared markers.

To practically apply CyTOFmerge, we recommend the following steps: *(1)* Collect the samples and divide them in two parts. *(2)* Design the first marker panel according to the biological question one wants to be answered. The marker panel would probably contain lineage markers, to differentiate between the major cell types, and cell state markers, for more detailed subtyping, and intracellular markers of interest[12] . *(3)* Stain the first part of the samples with the designed marker panel and measure the samples with CyTOF. *(4)* Apply the marker selection pipeline on the measured dataset using the first panel and obtain the most informative markers (i.e. shared markers). *(5)* Include those shared markers while designing the second panel of marker. *(6)* Add extra state or intracellular markers of interest to the second panel. *(7)* Stain the second part of the samples with the second marker panel and measure the samples with CyTOF. *(8)* Apply the imputation algorithm to all samples, combining both datasets from both panels, and create the imputed dataset in which each cell is represented by the unique markers from each panel (one of which is imputed), as well as the shared markers.

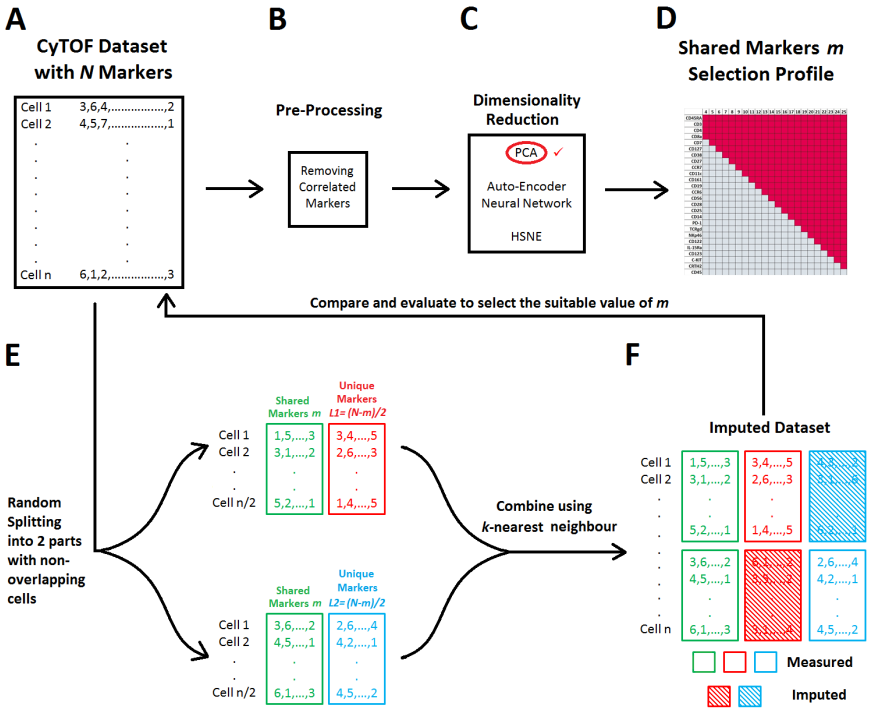
Importantly, we have shown that by combining panels a richer protein profile of cells can be acquired with which it becomes possible to find both abundant as well as rare cell populations. This opens possibilities to merge even more panels based on a common shared marker set as there is no fundamental limit to restrict to the combination of two panels.

BIBLIOGRAPHY

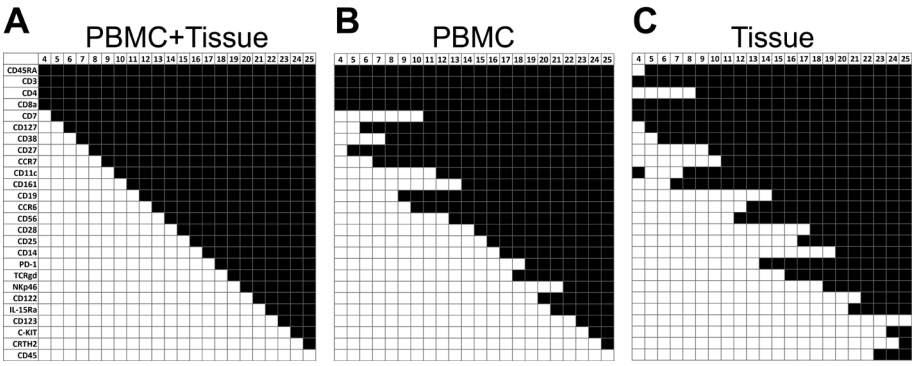
1. Bandura, D. R. *et al.* Mass cytometry: Technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* **81,** 6813–6822 (2009).
2. Spitzer, M. H. & Nolan, G. P. Mass Cytometry: Single Cells, Many Features. *Cell* **165,** 780–791 (2016).
3. Newell, E. W., Sigal, N., Bendall, S. C., Nolan, G. P. & Davis, M. M. Cytometry by Time-of-Flight Shows Combinatorial Cytokine Expression and Virus-Specific Cell Niches within a Continuum of CD8 + T Cell Phenotypes. *Immunity* **36,** 142–152 (2012).
4. Newell, E. W. *et al.* Combinatorial tetramer staining and mass cytometry analysis facilitate T-cell epitope mapping and characterization. *Nat. Biotechnol.* **31,** 623–629 (2013).
5. Amir, E. D. *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31,** 545–552 (2014).
6. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162,** 184–197 (2015).
7. Wong, M. T. *et al.* A High-Dimensional Atlas of Human T Cell Diversity Reveals Tissue-Specific Trafficking and Cytokine Signatures. *Immunity* **45,** 442–456 (2016).
8. van Unen, V. *et al.* Mass Cytometry of the Human Mucosal Immune System Identifies Tissue- and Disease-Associated Immune Subsets. *Immunity* **44,** 1227–1239 (2016).
9. Lavin, Y. *et al.* Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses. *Cell* **169,** 750–765 (2017).
10. Chevrier, S. *et al.* An Immune Atlas of Clear Cell Renal Cell Carcinoma. *Cell* **169,** 736–749 (2017).
11. Bendall, S. C., Nolan, G. P., Roederer, M. & Chattopadhyay, P. K. A deep profiler ’s guide to cytometry. *Trends Immunol.* **33,** 323–332 (2012).
12. Bendall, S. C. *et al.* Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science (80-.).* **332,** 687–696 (2011).
13. Pedreira, C. E. *et al.* Generation of Flow Cytometry Data Files with a Potentially Infinite Number of Dimensions. *Cytom. A* **73A,** (2008).
14. Lee, G., Finn, W. & Scott, C. Statistical file matching of flow cytometry data. *J. Biomed. Inform.* **44,** 663–676 (2011).
15. O’Neill, K. *et al.* Deep profiling of multitube flow cytometry data. *Bioinformatics* **31,** 1623–1631 (2015).
16. Costa, E. S. *et al.* Automated pattern-guided principal component analysis vs expert-based immunophenotypic classification of B-cell chronic lymphoproliferative disorders : a step forward in the standardization of clinical immunophenotyping. *Leukemia* **24,** 1927–1933 (2010).
17. van Dongen, J. *et al.* EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal , reactive and malignant leukocytes. *Leukemia* **26,** 1908–1975 (2012).
18. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34,** 1145–1160 (2016).
19. Samusik, N., Good, Z., Spitzer, M. H., Davis, K. L. & Nolan, G. P. Automated mapping of phenotype space with single-cell data. *Nat. Methods* **13,** 493–496 (2016).
20. Shlens, J. *A Tutorial on Principal Component Analysis*. (2005).
21. Hinton, G. & Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science (80-.).* **313,** 504–508 (2006).
22. Pezzotti, N., Höllt, T., Lelieveldt, B., Eisemann, E. & Vilanova, A. Hierarchical Stochastic Neighbor Embedding. in *Computer Graphics Forum (Proceedings of EuroVis 2016)* **35,** (2016).
23. Van Unen, V. *et al.* Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. *Nat. Commun.* **8,** 1–10 (2017).
24. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells

that Correlate with Prognosis. *Cell* **162,** 1–14 (2015).
25. Höllt, T. *et al.* Cytosplore : Interactive Immune Cell Phenotyping for Large Single-Cell Datasets. in *Computer Graphics Forum (Proceedings of EuroVis 2016)* **35,** (2016).
26. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn.* **9,** 2579–2605 (2008).
27. Pezzotti, N. *et al.* Approximated and User Steerable tSNE for Progressive Visual Analytics. *IEEE Trans. Vis. Comput. Graph.* **23,** 1739–1752 (2017).
28. Comaniciu, D. & Meer, P. Mean Shift : A Robust Approach Toward Feature Space Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **24,** 603–619 (2002).
29. Abdelaal, T. *et al.* Predicting Cell Populations in Single Cell Mass Cytometry Data. *Cytom. Part A* **95,** (2019).

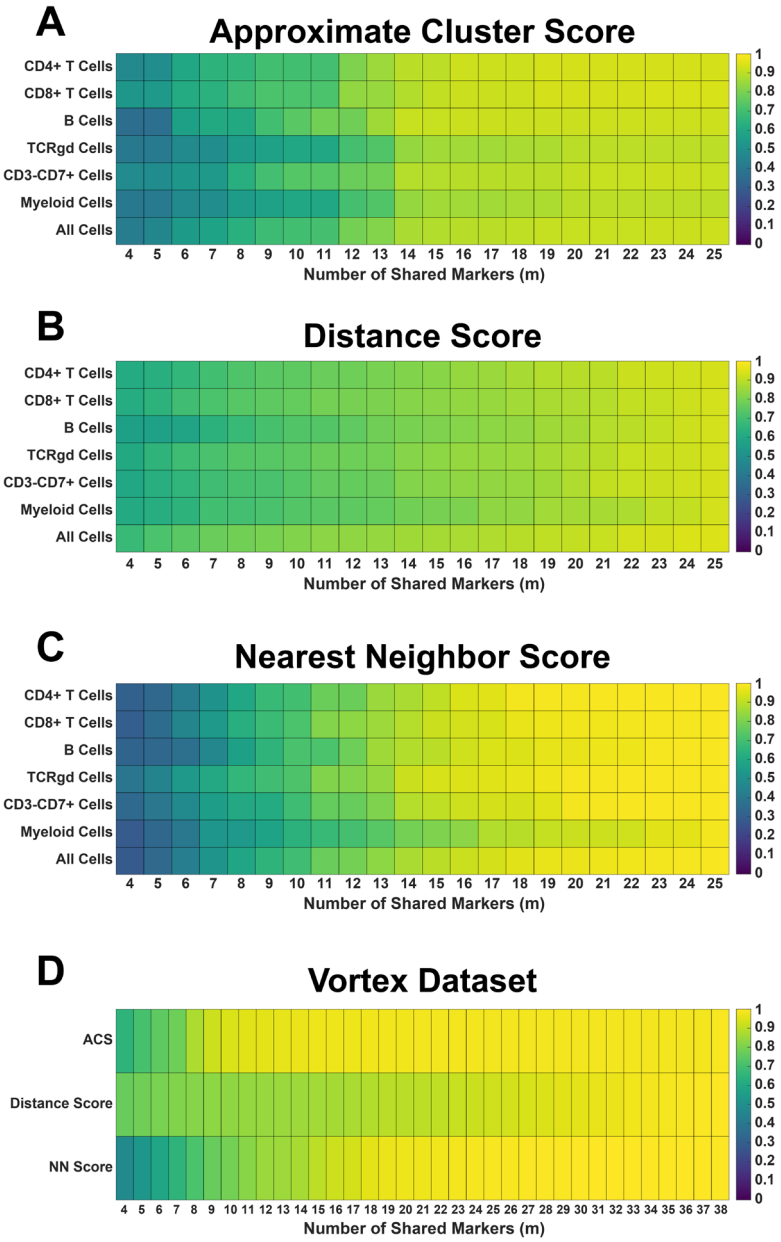
Supplementary Materials



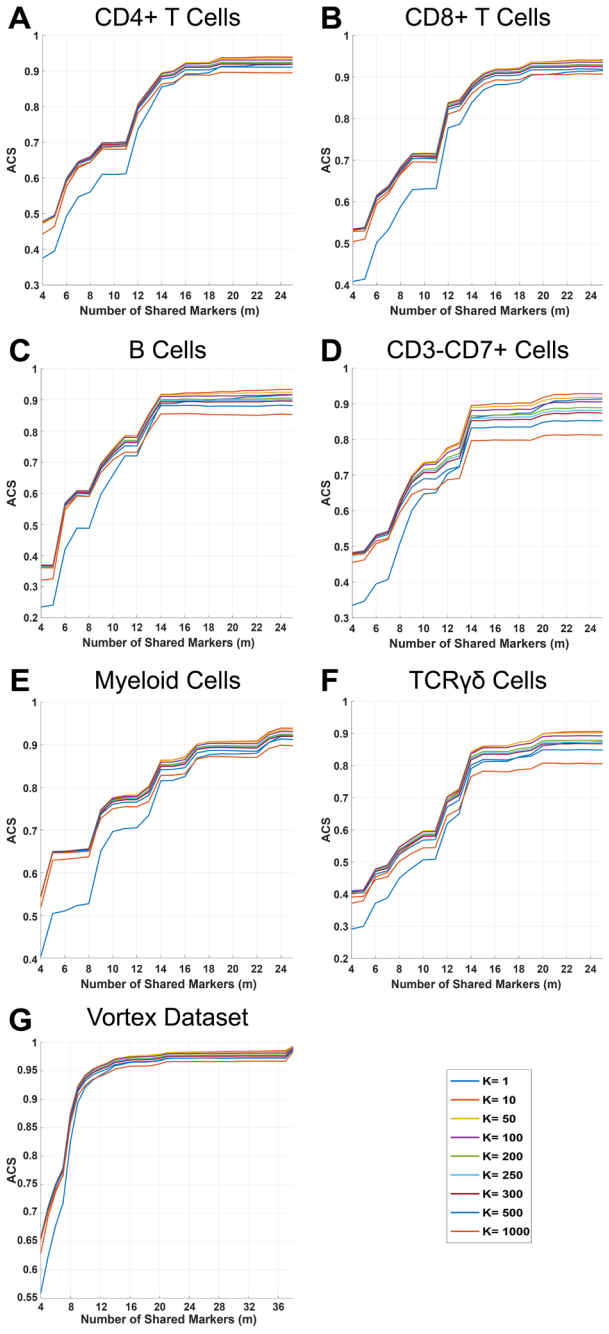
Supplementary Figure 7.1 Selection pipeline for shared markers (A) CyTOF dataset measured by a single panel, **(B)** calculate markers pairwise correlation and remove highly correlated markers, **(C)** apply dimensionality reduction to obtain the selection profile **(D)** for a wide range of *m* (red is selected, grey is not selected). **(E)** Evaluation of all values of *m* by randomly splitting the data, recombine it, and then comparing the imputed data **(F)** with the original data to select the minimal *m* with accepted performance.



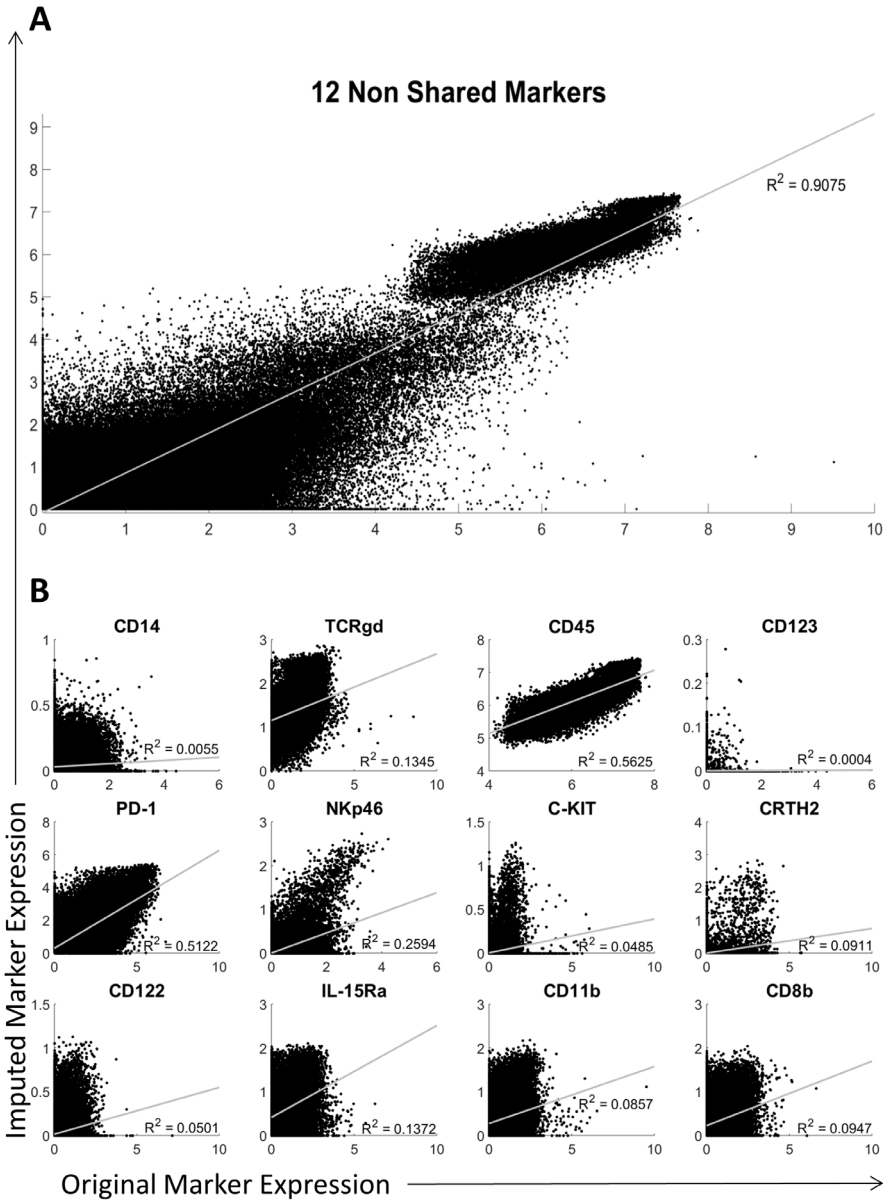
Supplementary Figure 7.2 Selected shared markers for PBMC and tissue: The PCA-based selected markers using **(A)** all samples (PBMC+Tissue), **(B)** using only PBMC samples and **(C)** using only tissue samples. (Marker ordering is based on the PCA selection profile, black is selected, white is not selected)



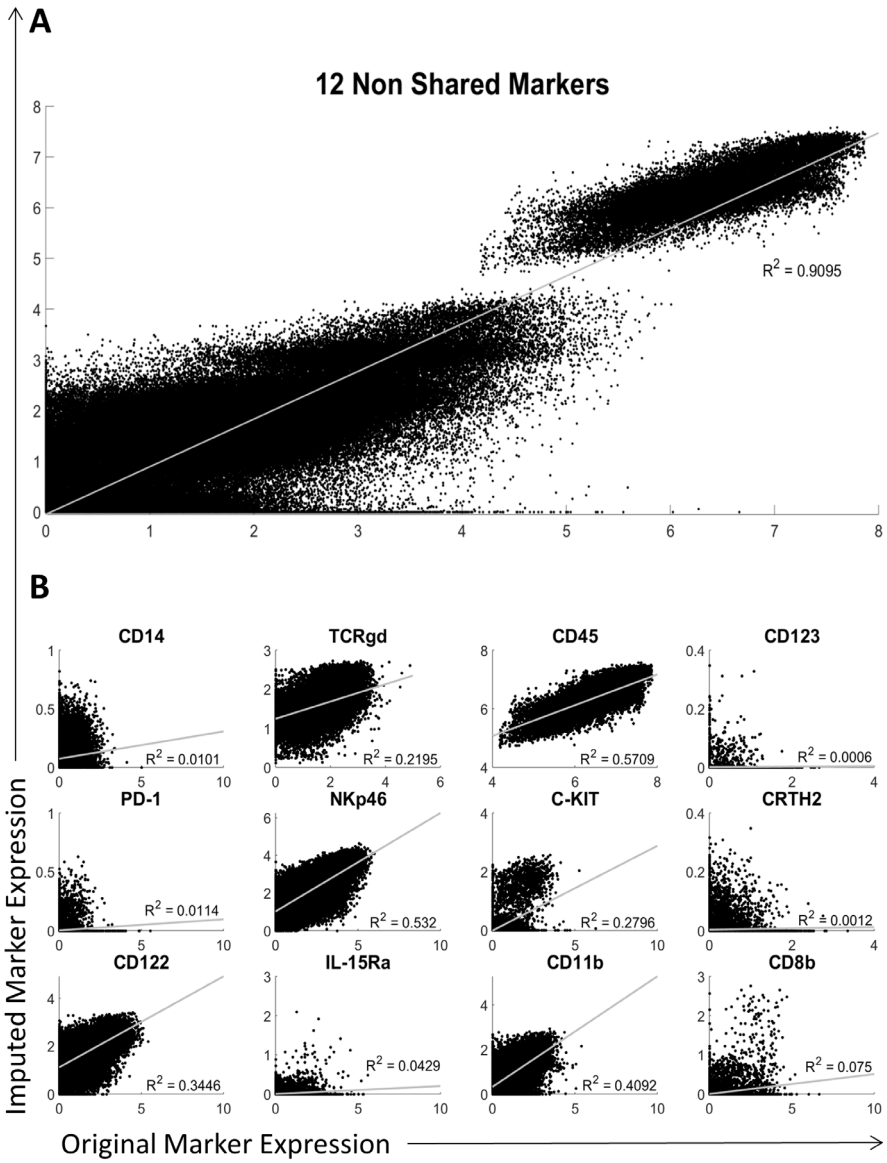
Supplementary Figure 7.3 Evaluation scores for imputed CyTOF data (A-C) Evaluation scores for the HMIS dataset as a function of shared markers for different lineages: **(A)** Approximate cluster score. **(B)** Distance score. **(C)** Nearest Neighbor score. These performance scores are calculated per lineage and for one dataset having all six lineages together (last row). **(D)** Evaluation scores for the Vortex dataset as a function of shared markers.



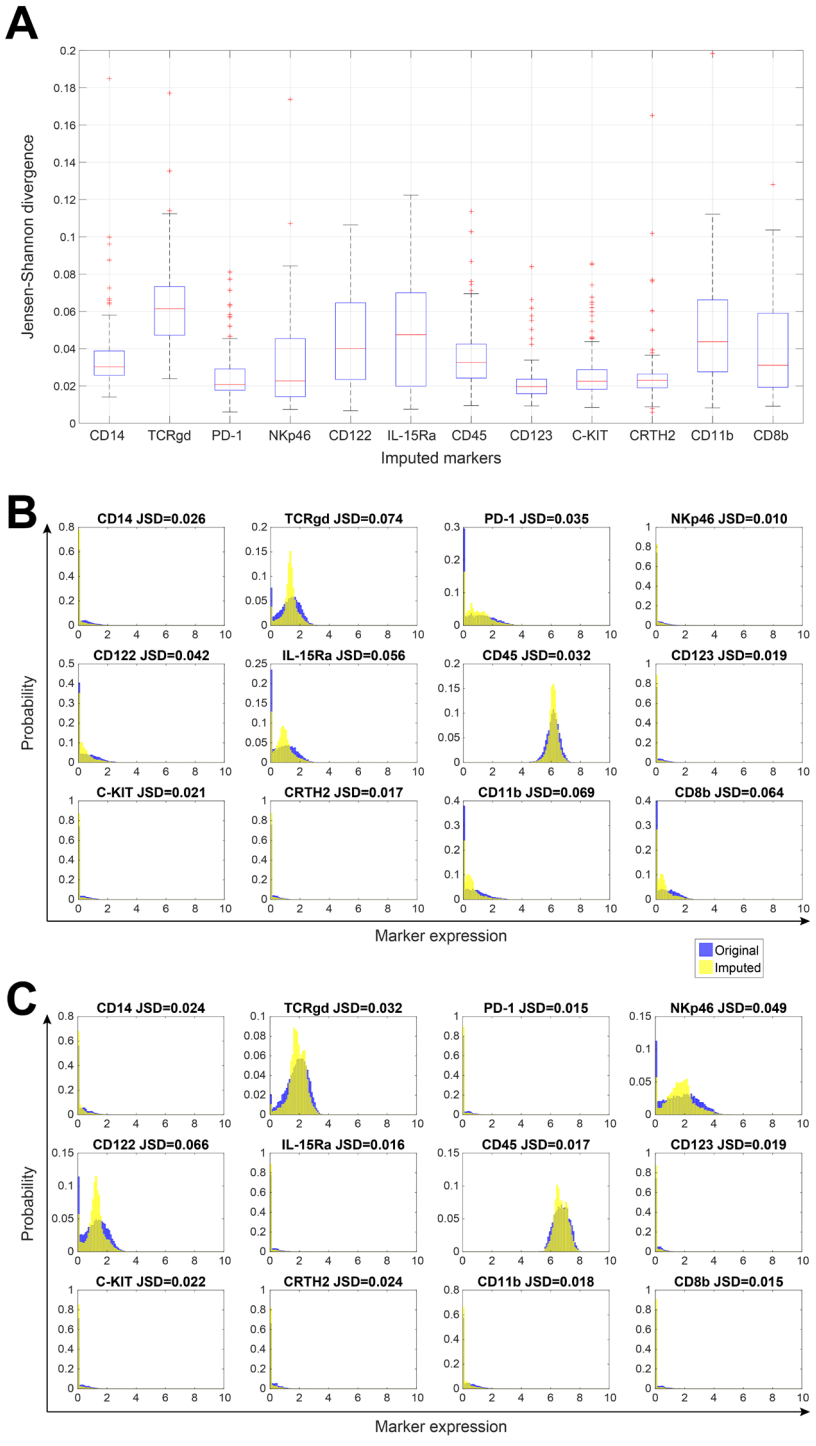
Supplementary Figure 7.4 *Approximate Cluster Score* vs the number of shared markers (*m*), combination was performed using the k-nearest neighbor algorithm for different values of *k* represented by the separate lines in each plot.



Supplementary Figure 7.5 Scatter plots showing the correlation between the original and the imputed expression values for the 12 non-shared markers of both panels A & B for the CD4+ T cells lineage: **(A)** all non-shared markers concatenated in one vector, showing a global high correlation, **(B)** separate scatter plots per marker, as shown within a specific lineage most of the markers are not expressed (≈ 0) resulting in a low correlation with the imputed values.

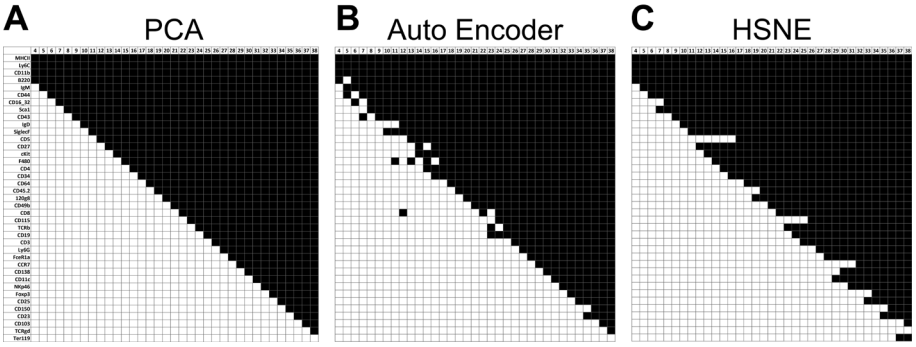


Supplementary Figure 7.6 Scatter plots showing the correlation between the original and the imputed expression values for the 12 non-shared markers of both panels A & B for the CD3-CD7+ cells lineage: **(A)** all non-shared markers concatenated in one vector, showing a global high correlation, **(B)** separate scatter plots per marker, as shown within a specific lineage most of the markers are not expressed (≈ 0) resulting in a low correlation with the imputed values.

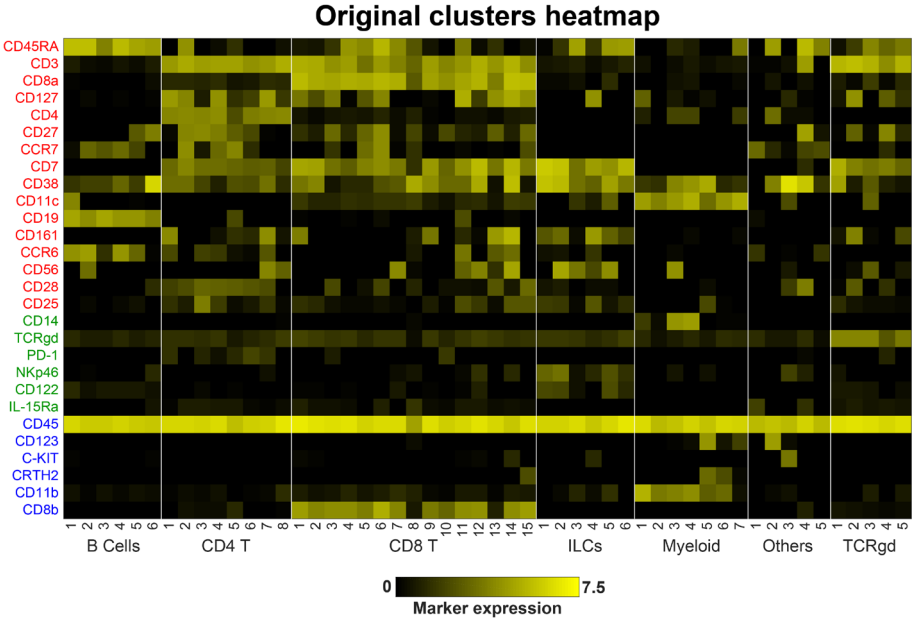


Supplementary Figure 7.7 (A) Box plots showing the Jensen-Shannon divergence (JSD) values for each of the 12 imputed markers across all 121 cell populations in the HMIS dataset. The JSD value measures the similarity between the original and the imputed distribution of one marker within one population. **(B-**

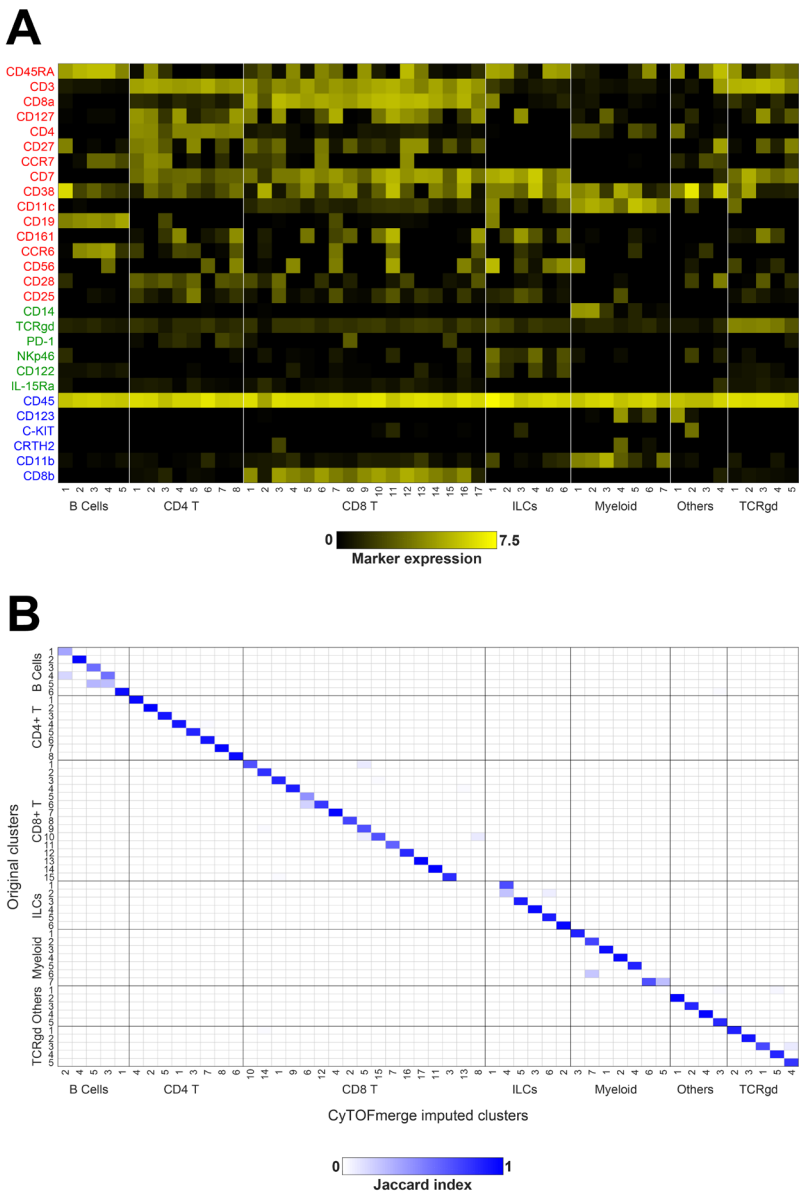
C) Histograms showing the original and the imputed distributions of the 12 imputed markers for **(B)** population CD4+ T cells 01, and **(C)** population CD3-CD7+ cells 01. For each marker, the JSD value is indicated.



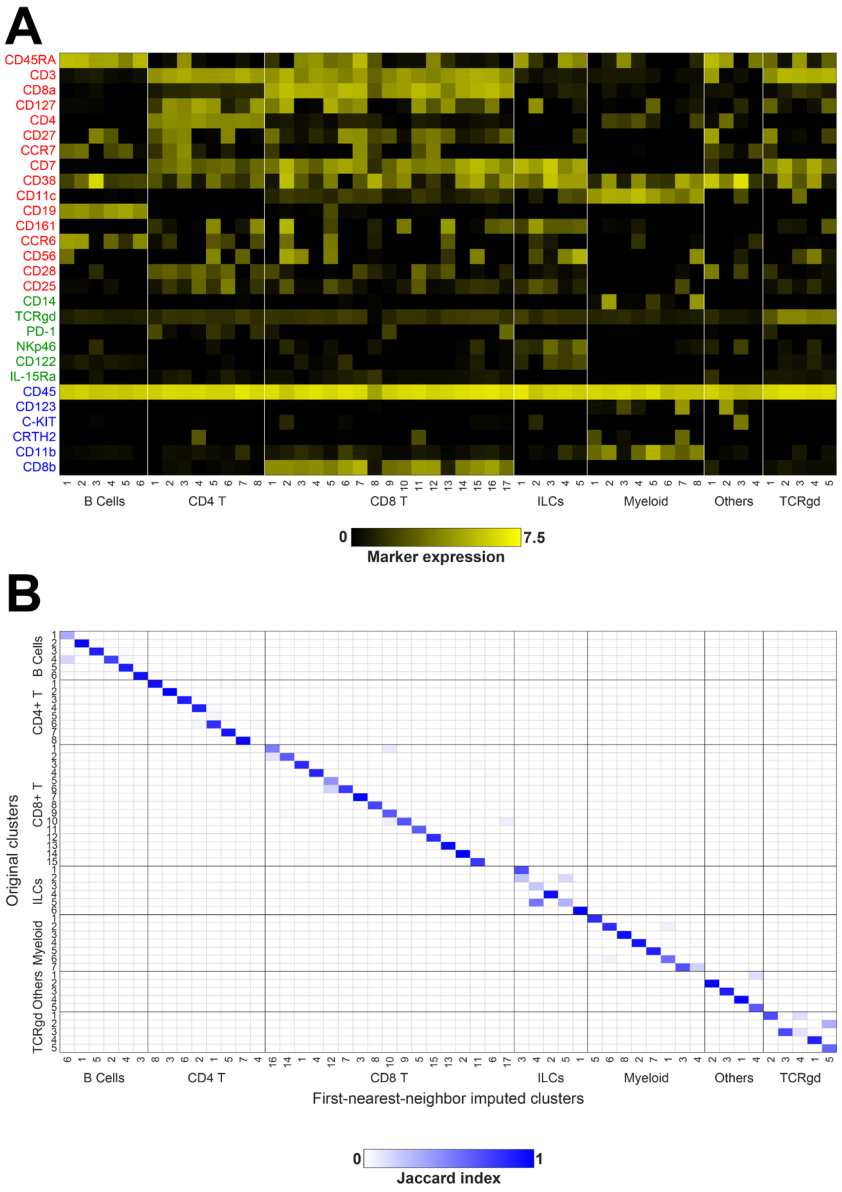
Supplementary Figure 7.8 Selected shared markers for the Vortex dataset. The selected markers that can best represent the dataset using **(A)** PCA, **(B)** Auto Encoder and **(C)** HSNE. (Marker ordering is based on the PCA selection profile, black is selected, white is not selected). No markers are removed during preprocessing.



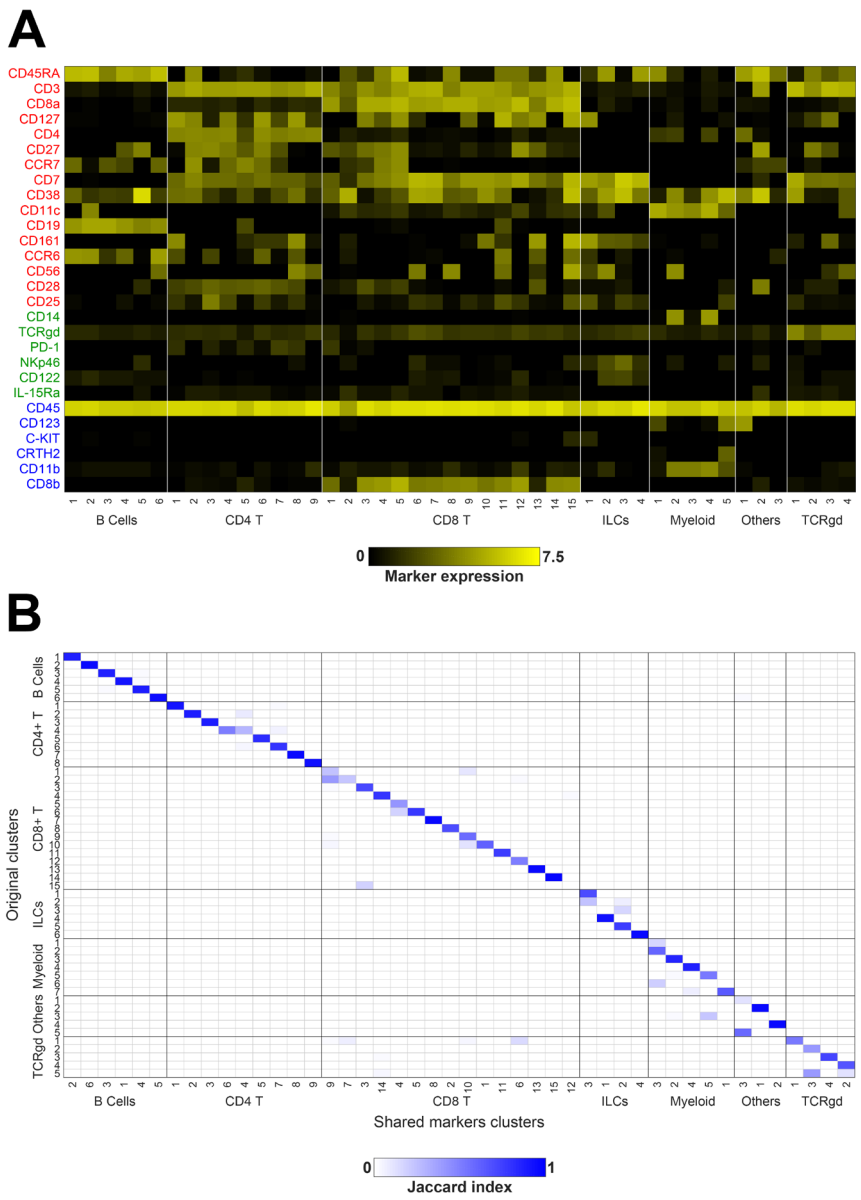
Supplementary Figure 7.9 HMIS original dataset clusters. Heatmap showing the median arcsinh5-transformed marker expression values (black-to-yellow scale) for the total 52 cell clusters obtained by clustering the original HMIS data using Phenograph. Marker colors indicate whether a marker is shared between panels or unique to a single panel, during panels combination (red is shared, green is unique to panel A, blue is unique to panel B).



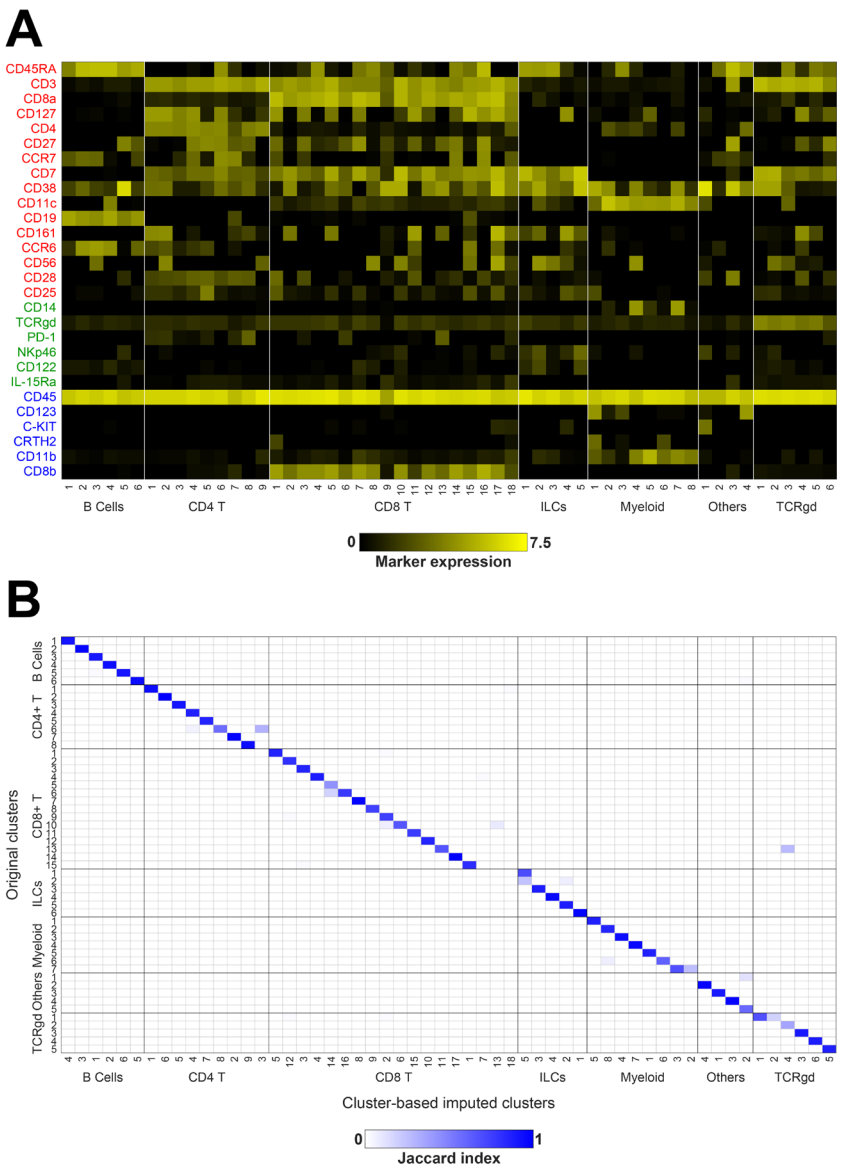
Supplementary Figure 7.10 HMIS imputed data by CyTOFmerge. (A) Heatmap showing the median arcsinh5-transformed marker expression values (black-to-yellow scale) for the total 52 cell clusters obtained by clustering the imputed HMIS data using Phenograph (with m = 16 and k = 50). Marker colors indicate whether a marker is shared between panels or unique to a single panel, during panels combination (red is shared, green is unique to panel A, blue is unique to panel B). **(B)** Pairwise Jaccard index map between the original and the imputed clusters of the HMIS dataset.



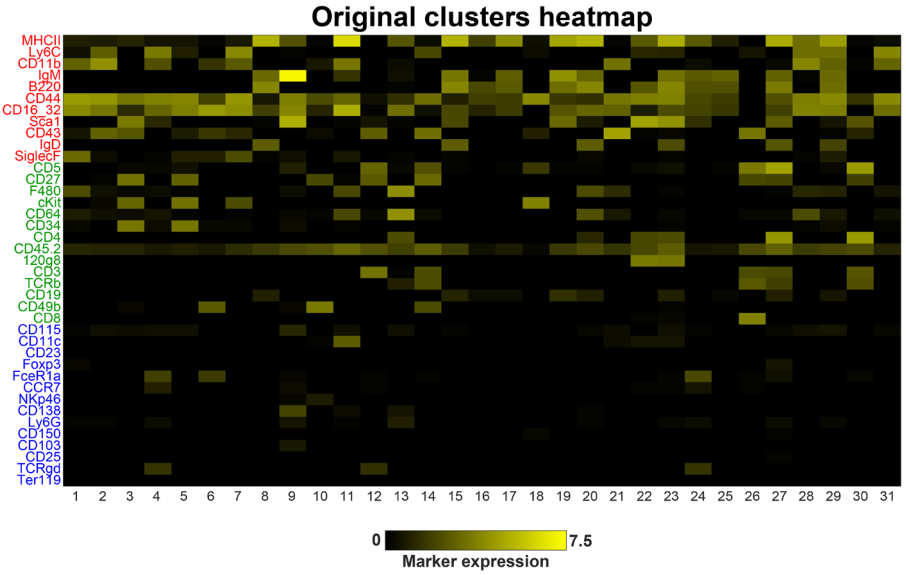
Supplementary Figure 7.11 HMIS imputed data by first-nearest-neighbor. (A) Heatmap showing the median arcsinh5-transformed marker expression values (black-to-yellow scale) for the total 53 cell clusters obtained by clustering the imputed HMIS data using Phenograph (with m = 16 and k = 1). Marker colors indicate whether a marker is shared between panels or unique to a single panel, during panels combination (red is shared, green is unique to panel A, blue is unique to panel B). **(B)** Pairwise Jaccard index map between the original and the imputed clusters of the HMIS dataset.



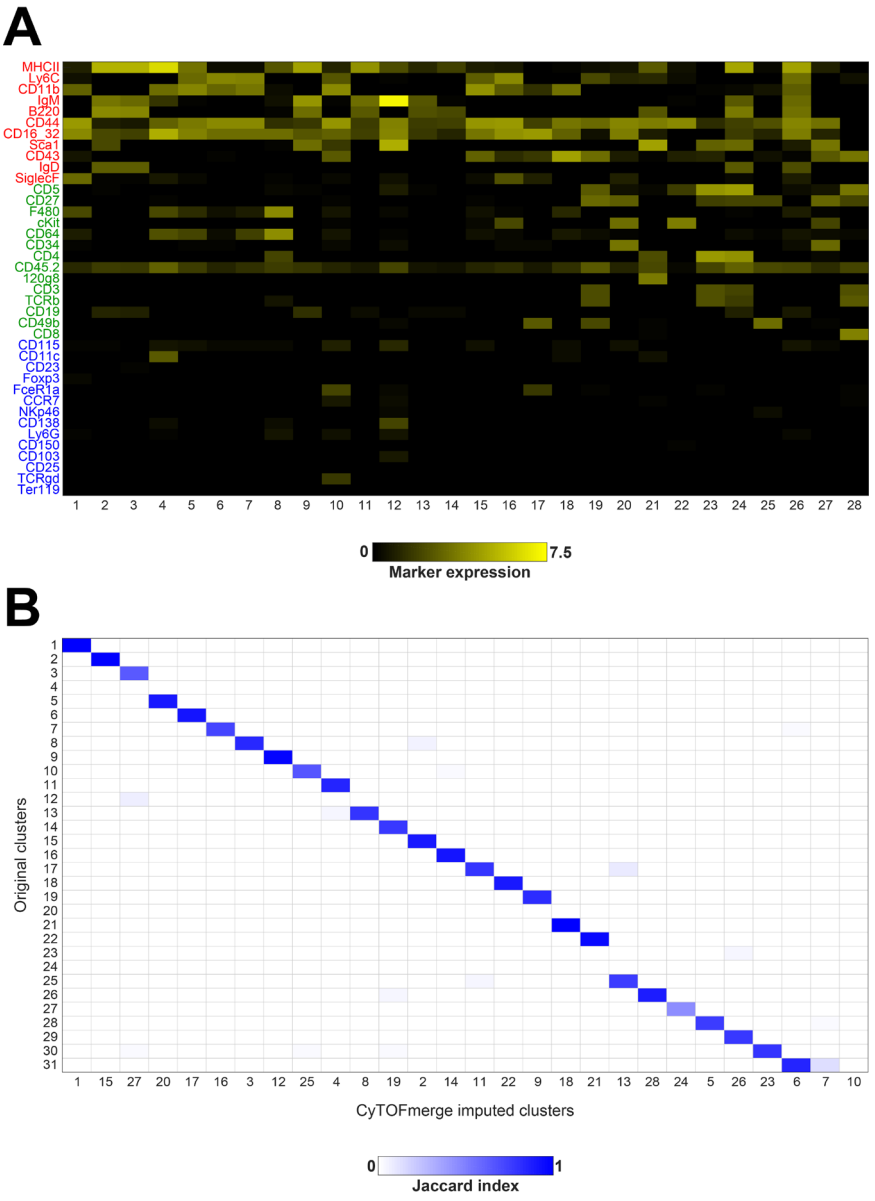
Supplementary Figure 7.12 HMIS shared markers clusters. (A) Heatmap showing the median arcsinh5-transformed marker expression values (black-to-yellow scale) for the total 42 cell clusters obtained by clustering the shared markers of the original HMIS data using Phenograph (m = 16). Marker colors indicate whether a marker is shared between panels or unique to a single panel, during panels combination (red is shared, green is unique to panel A, blue is unique to panel B). **(B)** Pairwise Jaccard index map between the original and the shared markers clusters of the HMIS dataset.



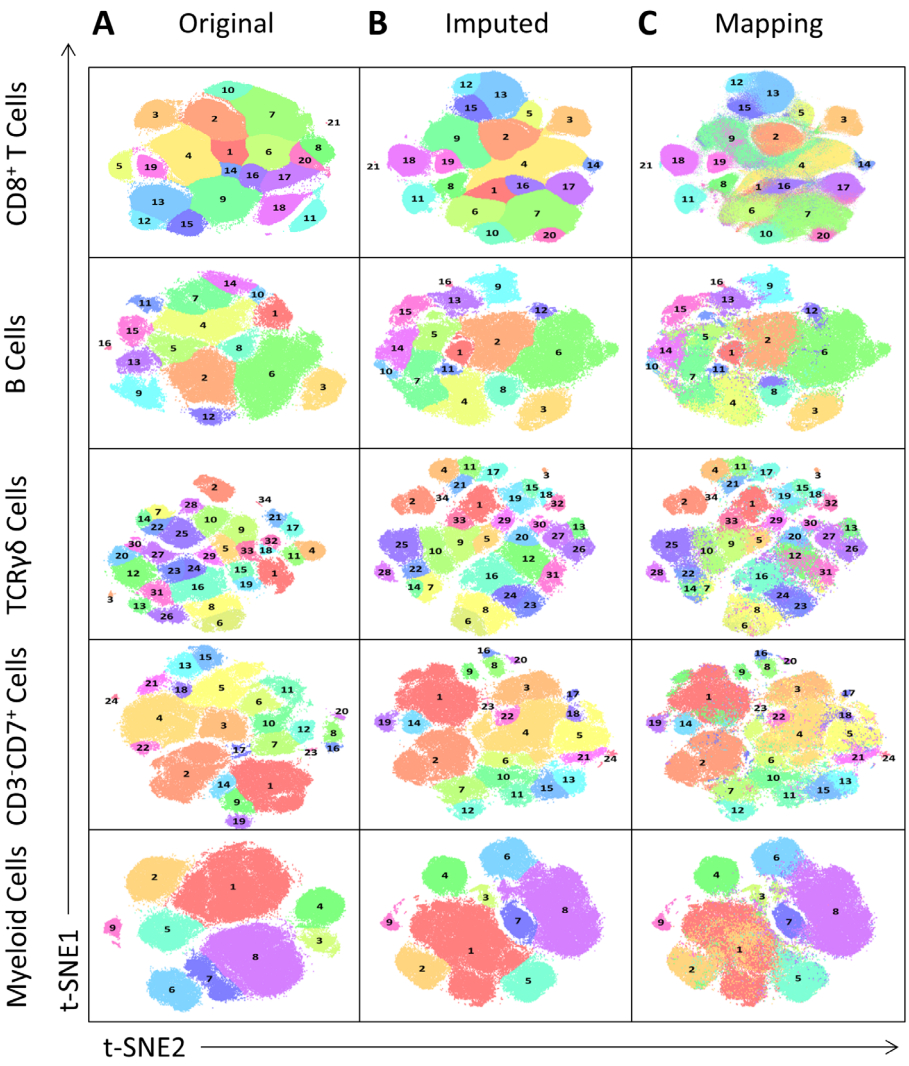
Supplementary Figure 7.13 HMIS imputed data by Cluster-based imputation. (A) Heatmap showing the median arcsinh5-transformed marker expression values (black-to-yellow scale) for the total 56 cell clusters obtained by clustering the imputed HMIS data using Phenograph (with m = 16 and k = 50, imputation performed within the same cluster found based on the shared markers space). Marker colors indicate whether a marker is shared between panels or unique to a single panel, during panels combination (red is shared, green is unique to panel A, blue is unique to panel B). **(B)** Pairwise Jaccard index map between the original and the imputed clusters of the HMIS dataset.



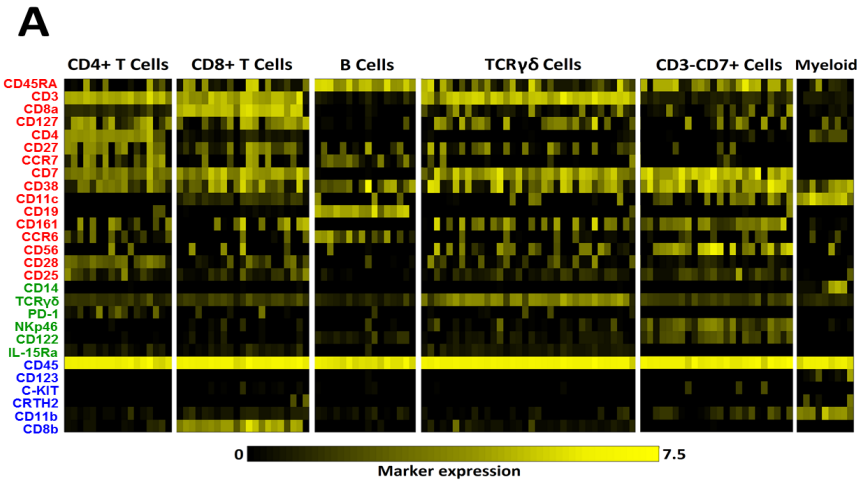
Supplementary Figure 7.14 Vortex original dataset clusters. Heatmap showing the median arcsinh5-transformed marker expression values (black-to-yellow scale) for the total 31 cell clusters obtained by clustering the original Vortex data using Phenograph. Marker colors indicate whether a marker is shared between panels or unique to a single panel, during panels combination (red is shared, green is unique to panel A, blue is unique to panel B).

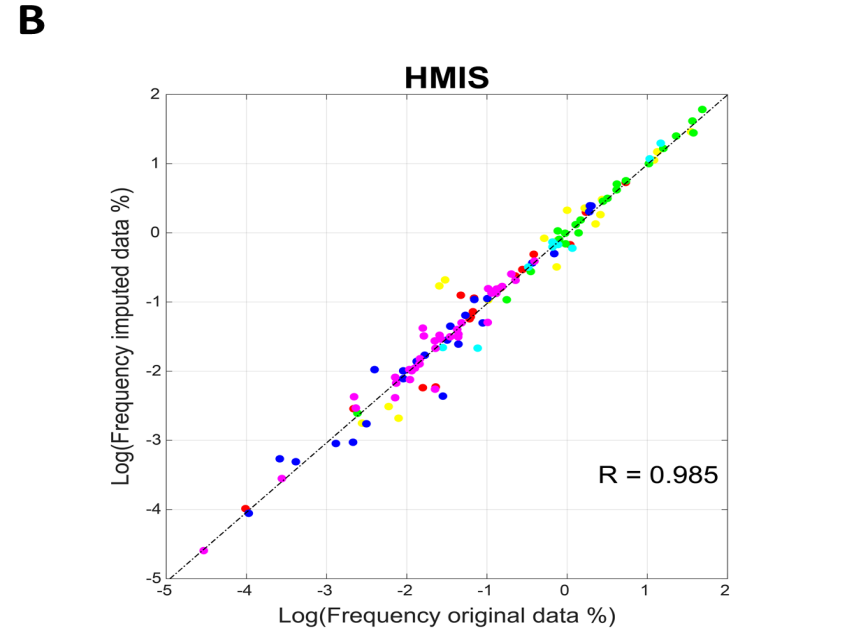


Supplementary Figure 7.15 Vortex imputed data by CyTOFmerge. (A) Heatmap showing the median arcsinh5-transformed marker expression values (black-to-yellow scale) for the total 28 cell clusters obtained by clustering the imputed Vortex data using Phenograph (with m = 11 and k = 50). Marker colors indicate whether a marker is shared between panels or unique to a single panel, during panels combination (red is shared, green is unique to panel A, blue is unique to panel B). **(B)** Pairwise Jaccard index map between the original and the imputed clusters of the Vortex dataset.

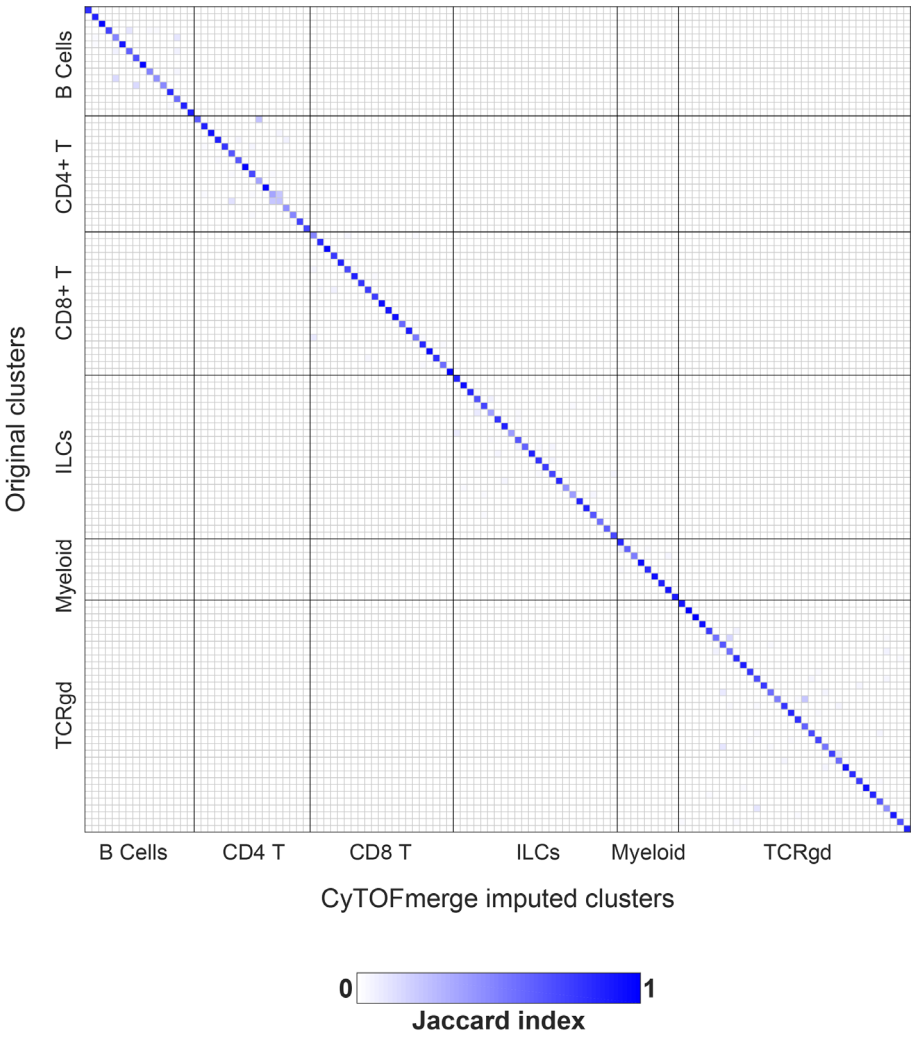


Supplementary Figure 7.16 Clustering of the original and the imputed datasets: t-SNE maps showing the different identified populations in each immune lineage, each row represent a separate lineage, column **(A)** shows the populations of the original data, column **(B)** shows the populations of the imputed data (for m=16, $L1$=6 and $L2$=6) and column **(C)** is the mapping of the original clusters labels on the t-SNE map of the imputed data.

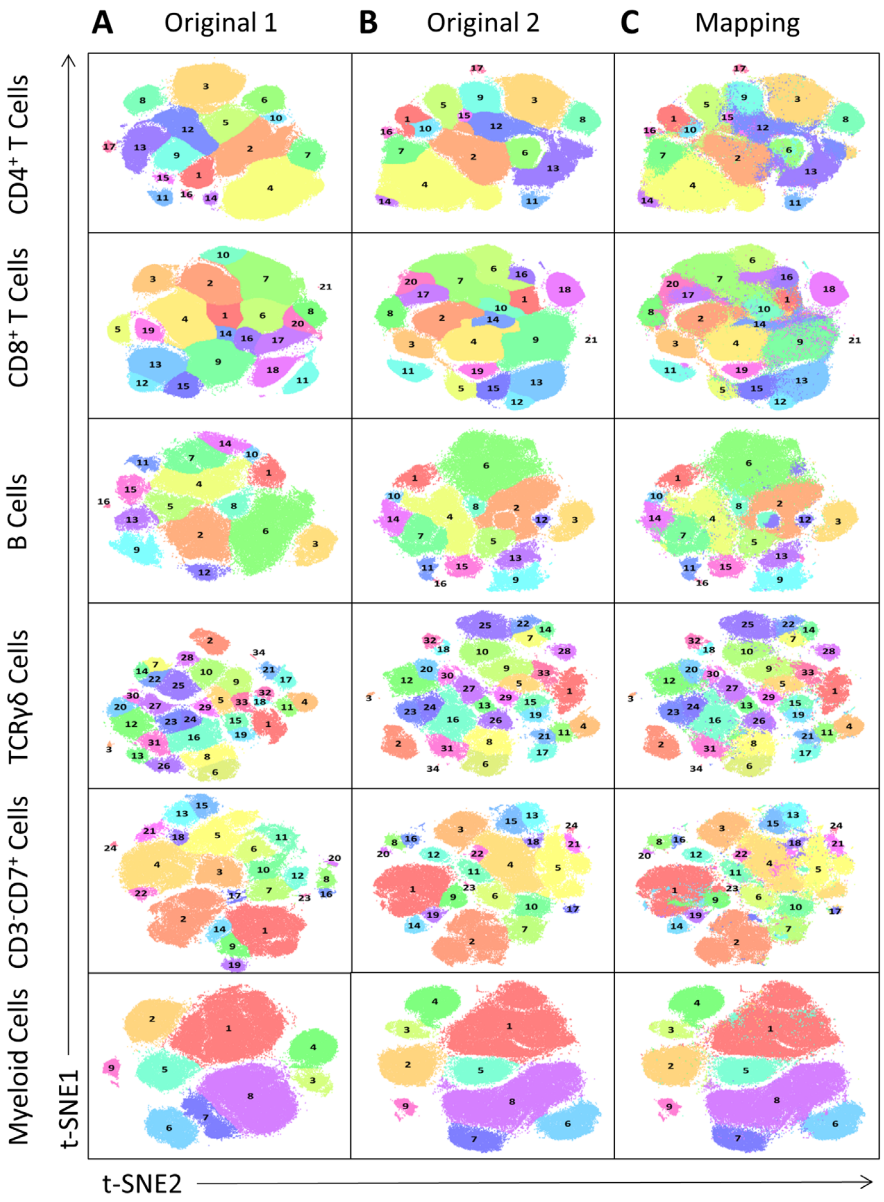




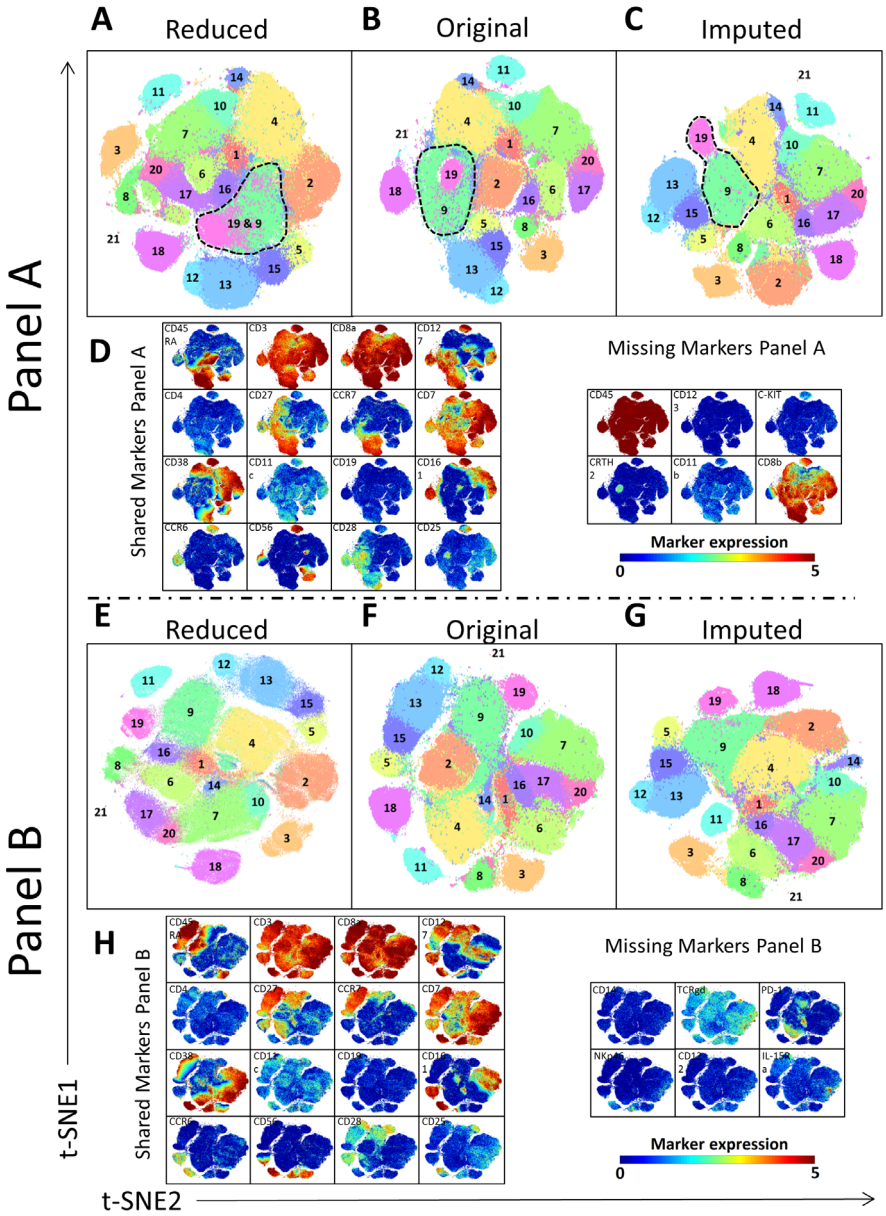
Supplementary Figure 7.17 (A) Heatmap of markers expression for the 121 characterized immune cells populations of the imputed dataset for *m* = 16. Black-to-yellow scale shows the median arcsinh-5 transformed values for the markers expression. Marker colors indicate whether a marker is shared between panels or unique to a single panel, during panels combination (red is shared, green is unique to panel A, blue is unique to panel B). **(B)** Scatter plots between original and imputed data population frequencies, the dashed line shows the least-squares fit error line, and the R value represents Pearson correlation coefficient between original and imputed frequencies.



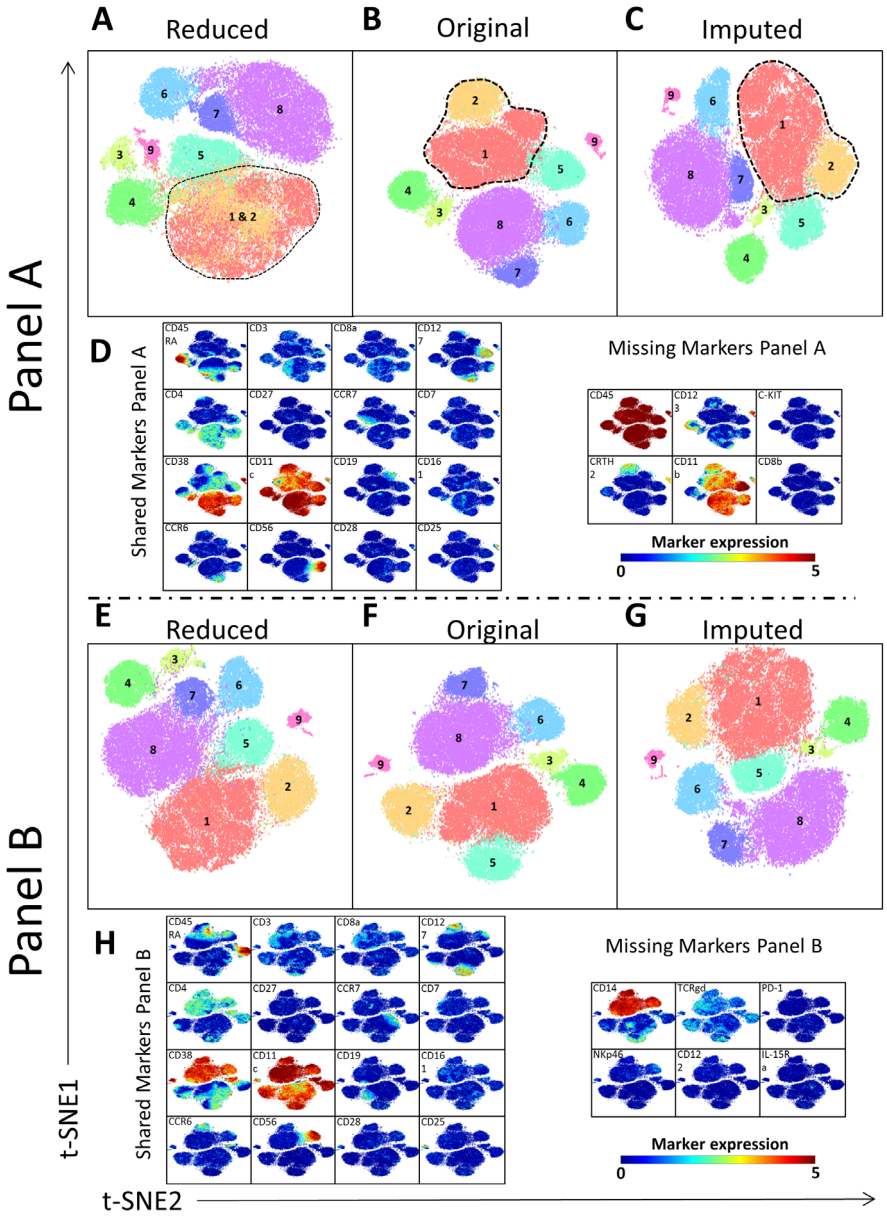
Supplementary Figure 7.18 Pairwise Jaccard index map between the original and the imputed clusters of the HMIS dataset, clustered using Cytosplore.



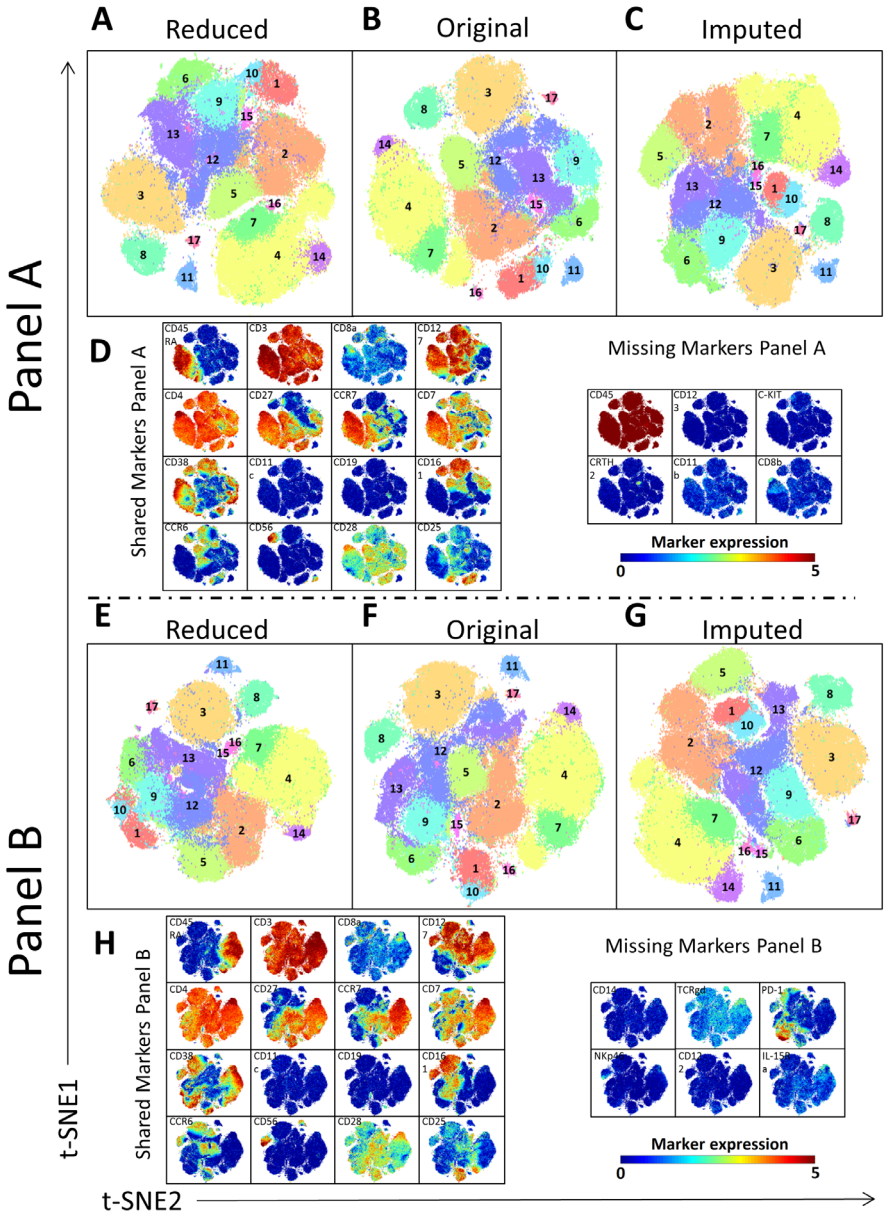
Supplementary Figure 7.19 Evaluation of t-SNE rerun: t-SNE maps showing the different identified populations in each immune lineage by running the t-SNE twice to the original data, each row represent a separate lineage, column **(A)** shows the populations of the original data for the first t-SNE map (Original 1), column **(B)** shows the populations of the original data for the second t-SNE map (Original 2) and column **(C)** is the mapping of the Original 1 clusters labels on the Original 2 t-SNE map.



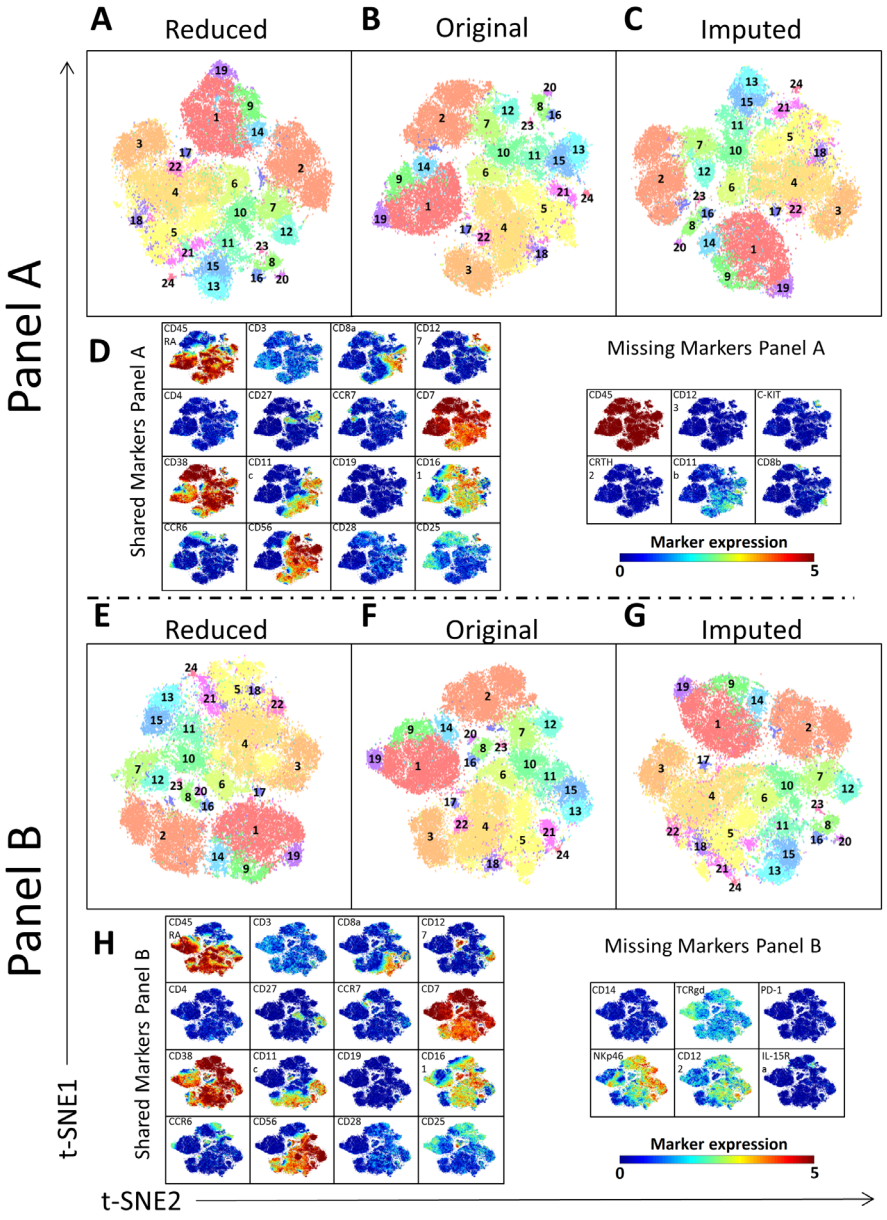
Supplementary Figure 7.20 Marker extension impact on identification of distinct populations in the CD8+ T Cells immune lineage: The upper half presents Panel A and the lower part presents Panel B. For each panel, the Reduced **(A, E)**, Original **(B, F)** and Imputed **(C, G)** t-SNE map are shown colored with the populations labels. Shared and missing markers expression profiles are shown on the Original t-SNE map **(D, H)**. The color scale bar shows the arcsinh-5 transformed values for the markers expression.



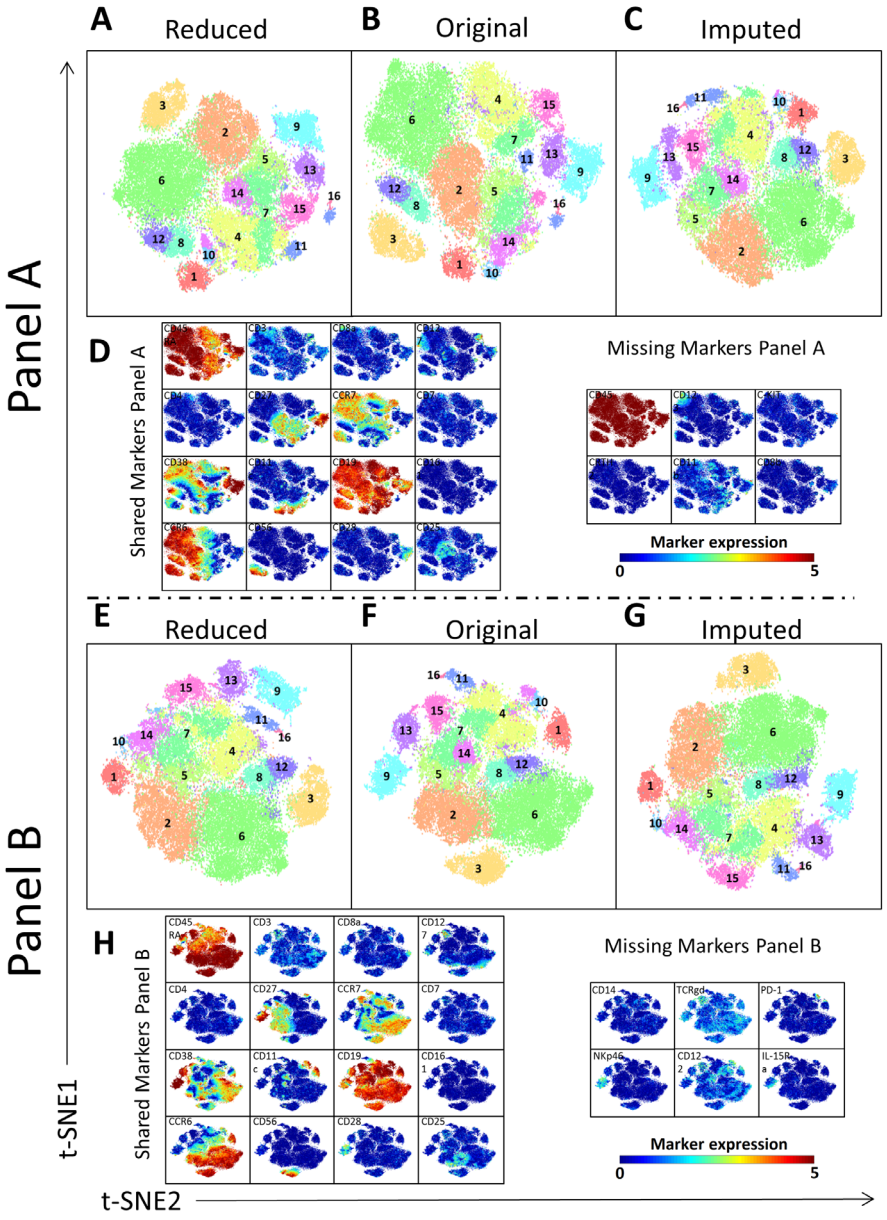
Supplementary Figure 7.21 Marker extension impact on identification of distinct populations in the Myeloid Cells immune lineage: The upper half presents Panel A and the lower part presents Panel B. For each panel, the Reduced **(A, E)**, Original **(B, F)** and Imputed **(C, G)** t-SNE map are shown colored with the populations labels. Shared and missing markers expression profiles are shown on the Original t-SNE map **(D, H)**. The color scale bar shows the arcsinh-5 transformed values for the markers expression.



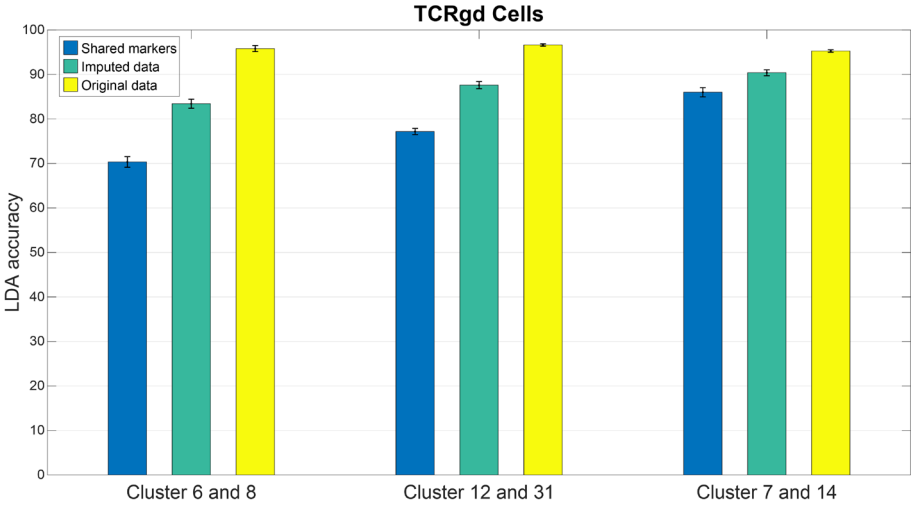
Supplementary Figure 7.22 Marker extension impact on identification of distinct populations in the CD4+ T Cells immune lineage: The upper half presents Panel A and the lower part presents Panel B. For each panel, the Reduced **(A, E)**, Original **(B, F)** and Imputed **(C, G)** t-SNE map are shown colored with the populations labels. Shared and missing markers expression profiles are shown on the Original t-SNE map **(D, H)**. The color scale bar shows the arcsinh-5 transformed values for the markers expression.



Supplementary Figure 7.23 Marker extension impact on identification of distinct populations in the CD3-CD7+ Cells immune lineage: The upper half presents Panel A and the lower part presents Panel B. For each panel, the Reduced **(A, E)**, Original **(B, F)** and Imputed **(C, G)** t-SNE map are shown colored with the populations labels. Shared and missing markers expression profiles are shown on the Original t-SNE map **(D, H)**. The color scale bar shows the arcsinh-5 transformed values for the markers expression.



Supplementary Figure 7.24 Marker extension impact on identification of distinct populations in the B Cells immune lineage: The upper half presents Panel A and the lower part presents Panel B. For each panel, the Reduced **(A, E)**, Original **(B, F)** and Imputed **(C, G)** t-SNE map are shown colored with the populations labels. Shared and missing markers expression profiles are shown on the Original t-SNE map **(D, H)**. The color scale bar shows the arcsinh-5 transformed values for the markers expression.



Supplementary Figure 7.25 LDA classification accuracy for clusters 6-8, 12-3 and 7-14, in the TCRγδ cells from the HMIS dataset. Classification is applied using the 16 shared markers only, all 28 markers from the imputed dataset, and all 28 markers from the original dataset. Error bar shows the performance variation across the 5-folds of the cross validation.

CHAPTER 8

SPAGE: SPATIAL GENE ENHANCEMENT USING SCRNA-SEQ

Tamim Abdelaal

Soufiane Mourragui

Marcel J.T. Reinders

Ahmed Mahfouz

This chapter is published in: *Nucleic Acid Research* (2020) 48(18): e107, doi: 10.1093/nar/gkaa740.
Supplementary material is available online at: https://academic.oup.com/nar/article/48/18/e107/5909530#208175714

Single-cell technologies are emerging fast due to their ability to unravel the heterogeneity of biological systems. While scRNA-seq is a powerful tool that measures whole-transcriptome expression of single cells, it lacks their spatial localization. Novel spatial transcriptomics methods do retain cells spatial information but some methods can only measure tens to hundreds of transcripts. To resolve this discrepancy, we developed SpaGE, a method that integrates spatial and scRNA-seq datasets to predict whole-transcriptome expressions in their spatial configuration. Using five dataset-pairs, SpaGE outperformed previously published methods and showed scalability to large datasets. Moreover, SpaGE predicted new spatial gene patterns that are confirmed independently using in situ hybridization data from the Allen Mouse Brain Atlas.

8.1 INTRODUCTION

Single cell technologies rapidly developed over the last decade and have become valuable tools for enhancing our understanding of biological systems. Single-cell RNA-sequencing (scRNA-seq) allows unbiased measurement of the entire gene expression profile of each individual cell and has become the de facto technology used to characterize the cellular composition of complex tissues[1,2]. However, single cells often have to be dissociated before performing scRNA-seq and results in losing the spatial context and hence limits our understanding of cell identities and relationships. Recently, spatial transcriptomics technologies have advanced and provide localizations of gene expressions and cellular structure at the cellular level[3,4]. Many current protocols can be divided in two categories: 1) imaging-based methods (e.g. osmFISH, MERFISH and seqFISH+)[5–7], and 2) sequencing-based methods (e.g. STARmap and Slide-seq)[8,9]. Imaging-based protocols have a high gene detection sensitivity; capturing high proportion of the mRNA molecules with relatively small dropout rate. While seqFISH+ and the latest generation of MERFISH can measure up to ~10,000 genes[7,10], many different imaging-based protocols are often limited in the number of genes that can be measured simultaneously. On the other hand, sequencing-based protocols like STARmap can scale up to thousands of genes, it has a relatively lower gene detection sensitivity. Slide-seq is not limited in the number of measured genes and can be used to measure the whole transcriptome. However, similar to STARmap, Slide-seq suffers from a low gene detection sensitivity. In addition, osmFISH, MERFISH and STARmap can capture genes at the single-molecule resolution, which can be averaged or aggregated to the single-cell level. While Slide-seq has a resolution of 10μm, which is comparable to the average cell size, but does not always represent a single-cell.

Given the complementary information provided by both scRNA-seq and spatial transcriptomics data, integrating both types would provide a more complete overview of cell identities and interactions within complex tissues. This integration can be performed in two different ways[11]: 1) dissociated single-cells measured with scRNA-seq can be mapped to their physical locations in the tissue[12–14], or 2) missing gene expression measurements in the spatial data can be predicted from scRNA-seq. In this study, we focus on the second challenge in which measured gene expressions of spatial cells can be enhanced by predicting the expression of unmeasured genes based on scRNA-seq data of a matching tissue. Several methods have addressed this problem using various data integration approaches to account for the differences between the two data types[15–18]. All these methods rely on joint dimensionality reduction methods to embed both spatial and scRNA-seq data into a common latent space. For example, Seurat uses canonical correlation analysis (CCA), Liger uses non-negative matrix factorization (NMF), and Harmony uses principal component analysis (PCA). While Seurat, Liger and Harmony rely on linear methods to embed the data, gimVI uses a non-linear deep generative model. Despite recent benchmarking efforts[19], a comprehensive

evaluation of these methods for the task of spatial gene prediction from dissociated cells is currently lacking. For example, Seurat, Liger and gimVI, have only been tested using relatively small datasets (<2,000 cells)[15,16,18]. It is thus not clear whether a complex model, like gimVI, is really necessary. Moreover, Seurat, Harmony and gimVI lack interpretability of the integration procedure, so that it does not become clear which genes contribute in the prediction task.

Here, we present SpaGE (Spatial Gene Enhancement), a robust, scalable and interpretable machine-learning method to predict unmeasured genes of each cell in spatial transcriptomic data through integration with scRNA-seq data from the same tissue. SpaGE relies on domain adaptation using PRECISE[20] to correct for differences in sensitivity of transcript detection between both single-cell technologies, followed by a k-nearest-neighbor (kNN) prediction of new spatial gene expression. We demonstrate that SpaGE outperforms state-of-the-art methods by accurately predicting unmeasured gene expression profiles across a variety of spatial and scRNA-seq dataset pairs of different regions in the mouse brain. These datasets include a large spatial data with more than 60,000 cells, used to illustrate the scalability and computational efficiency of SpaGE compared to other methods.

8.2 MATERIALS AND METHODS

8.2.1 SPAGE ALGORITHM

The SpaGE algorithm takes as input two gene expression matrices corresponding to the scRNA-seq data (reference) and the spatial transcriptomics data (query). Based on the set of shared genes between the two datasets, SpaGE enriches the spatial transcriptomics data using the scRNA-seq data, by predicting the expression of spatially unmeasured genes. The SpaGE algorithm can be divided in two major steps: (i) Alignment of the two datasets using the domain adaptation algorithm PRECISE[20], and (ii) gene expression prediction using k-nearest-neighbor regression.

First, PRECISE was used to project both datasets into a common latent space. Let $R_{(n \times g)}$ be the gene expression matrix of the reference dataset having n cells and g genes, and let $Q_{(m \times h)}$ be the gene expression matrix of the query dataset having m cells and h genes. Using the set of shared genes $p = g \cap h$, PRECISE applies independent Principal Component Analysis (PCA) for each dataset to define two independent sets of Principal Components (*PCs*), such that:

$$R_{(n \times p)} = R'_{(n \times d)} PC_{r\,(d \times p)} \quad \text{with } PC_r PC_r^T = I_d \tag{8.1}$$

and

$$Q_{(m \times p)} = Q'_{(m \times d)} PC_{q\,(d \times p)} \quad \text{with } PC_q PC_q^T = I_d \tag{8.2}$$

where d is the number of desired *PCs*, PC_r and PC_q represents the principal components of the reference and the query datasets, respectively. We choose d = 50 for the

STARmap_AllenVISp, **MERFISH_Moffit** and **seqFISH_AllenVISp** dataset pairs, and d = 30 for all the **osmFISH** dataset pairs. Next, PRECISE compares these independent *PCs* by computing the cosine similarity matrix and decomposing it by SVD[21]:

$$PC_r {PC_q}^T = U\Sigma V^T \tag{8.3}$$

where *U* and *V* represent orthogonal (of size d) transformations on the reference and query *PCs*, respectively, and Σ is a diagonal matrix. *U* and *V* are then used to align the *PCs*, yielding the so-called Principal Vectors (*PVs*), such that:

$$PV_r = U^T PC_r \tag{8.4}$$

and

$$PV_q = V^T PC_q \tag{8.5}$$

PV_r and PV_q are the principal vectors of the reference and the query datasets, respectively, retaining the same information as the principal components. However, these *PVs* have now a one-to-one correspondence as their cosine similarity matrix is diagonal (the matrix Σ). *PVs* are pairs of vectors $\left(PV_r^1, PV_q^1\right), \ldots, \left(PV_r^d, PV_q^d\right)$ sorted in decreasing order based of similarity. To remove noisy components, we choose a limited number of *PVs*, d', for further analysis, where the cosine similarity is higher than a certain threshold (0.3). The reference *PVs*, PV_r, are then used to project and align both the scRNA-seq (reference) and the spatial transcriptomics (query) datasets:

$$R_{aligned\ (n \times d')} = \ R_{(n \times p)} PV_{r\ (p \times d')}^T \tag{8.6}$$

and

$$Q_{aligned\ \ (m \times d')} = \ Q_{(m \times p)} PV_{r\ (p \times d')}^T \tag{8.7}$$

After aligning the datasets, SpaGE predicts the expression of the spatially unmeasured genes, $l = g - p$, from the scRNA-seq dataset. For each spatial cell $i \in m$, we define the *k*-nearest-neighbors (*k* = 50) from the n dissociated scRNA-seq cells, using the cosine distance. Next, we calculate an array of weights w_{ij} between spatial cell i and its nearest neighbors $j \in NN(i)$. Out of the 50 neighbors, we only keep neighbors with positive cosine similarity with cell i (i.e. cosine distance < 1), such that:

$$\forall\ j\ \in NN(i)\ and\ dist(i,j) < 1$$

$$w_{ij} = 1 - \frac{dist(i,j)}{\sum_j dist(i,j)} \tag{8.8}$$

$$w_{ij} = \frac{w_{ij}}{length(w_{ij}) - 1} \tag{8.9}$$

The predicted expression Y_{il} of the set of spatially unmeasured genes l for cell i is calculated as a weighted average of the nearest neighbors dissociated cells:

$$Y_{il} = \sum_{\substack{j \in NN(i) \\ dist(i,j)<1}} w_{ij} * R_{jl} \tag{8.10}$$

8.2.2 GENE CONTRIBUTION TO THE INTEGRATION

To evaluate the contribution of each gene in forming this common latent space PV_r, we calculated the gene contribution C_g of gene g as follows:

$$C_g = \sum_{i=1}^{d} \beta_{gi}^2 \tag{8.11}$$

where β_{gi} is the loading of gene g to the i-th principal vector in PV_r, and d' is the final number of *PVs* in PV_r. To obtain the top contributing genes, the C_g values are sorted in descending order across all genes. We used the same criteria to calculate the contribution of each gene for dataset-specific *PCs* or *PVs*.

8.2.3 DATASETS

We used six dataset pairs (Table 8.1) composed of four scRNA-seq datasets (**AllenVISp**[22], **AllenSSp**[23], **Zeisel**[24] and **Moffit**[4]) and four spatial transcriptomics datasets (**STARmap**[8], **osmFISH**[5], **MERFISH**[4] and **seqFISH+**[7]). The **AllenVISp** (GSE115746) and the **AllenSSp** datasets were downloaded from https://portal.brain-map.org/atlases-and-data/rnaseq. The **AllenVISp** is obtained from the 'Cell Diversity in the Mouse Cortex – 2018' release. The **AllenSSp** is obtained from the 'Cell Diversity in the Mouse Cortex and Hippocampus' release of October 2019. We downloaded the whole dataset and used the metadata to only select cells from the SSp region. The **Zeisel** dataset (GSE60361) was downloaded from http://linnarssonlab.org/cortex/, while the **Moffit** 10X dataset (GSE113576) was downloaded from GEO.

Table 8.1 Summary of the dataset pairs used in this study

Spatial_scRNA-seq dataset pair	Spatial data			scRNA-seq data		
	# of cells	# of genes	Tissue	# of cells	# of genes	Tissue
STARmap_AllenVISp[8,22]	1,549	1,020	VISc	14,249	34,617	VISc
osmFISH_Zeisel[5,24]	3,405	33	SMSc	1,691	15,075	SMSc
osmFISH_AllenSSp[5,23]	3,405	33	SMSc	5,577	30,527	SMSc
osmFISH_AllenVISp[5,22]	3,405	33	SMSc	14,249	34,617	VISc
MERFISH_Moffit[4]	64,373	155	POR	31,299	18,646	POR
seqFISH_AllenVISp[7,22]	524	10,000	Cortex	14,249	34,617	VISc

VISc: Visual cortex; SMSc: Somatosensory cortex; POR: Pre-optic region

The **STARmap** dataset was downloaded from the STARmap resources website (https://www.starmapresources.com/data). We obtained the gene count matrix and the cell position information for the largest 1020-gene replicate. Cell locations and morphologies were identified using Python code provided by the original study (https://github.com/weallen/STARmap). The **osmFISH** dataset was downloaded as loom file from http://linnarssonlab.org/osmFISH/, we obtained the gene count matrix and the metadata using the loompy Python package. The **MERFISH** dataset was downloaded from

Dryad repository (https://doi.org/10.5061/dryad.8t8s248), we used the first naïve female mouse (Animal_ID = 1). The **seqFISH+** dataset was obtained from the seqFISH-PLUS GitHub repository (https://github.com/CaiGroup/seqFISH-PLUS), we used the gene count matrix of the mouse cortex dataset.

8.2.4 DATA PREPROCESSING

For all the scRNA-seq datasets, we filtered out genes expressed in less than 10 cells. No filtration was applied on the cells, except for the **AllenVISp** dataset for which we filtered low quality cells provided from the metadata ('Low Quality' and 'No Class' cells). For the **Zeisel** dataset, we only used the somatosensory cortex cells excluding the hippocampus cells. Next, scRNA-seq datasets were normalized by dividing the counts within each cell by the total number of transcripts within that cell, scaling by 10^6 and log1p transformed. Further, we scaled the data by making each gene centered and scaled (zero mean and unit variance) using the SciPy Python package[25].

For spatial transcriptomics datasets all gene were used, except for the **MERFISH** dataset for which we removed the blanks genes and the *Fos* gene (non-numerical values). Additionally, we filtered out cells labeled as 'Ambiguous' from the **MERFISH** dataset. Similar to the **Zeisel** dataset, we only kept cells from cortical regions for the **osmFISH** dataset ('Layer 2-3 lateral', 'Layer 2-3 medial', 'Layer 3-4', 'Layer 4','Layer 5', 'Layer 6' and 'Pia Layer 1'). For the **seqFISH+** dataset, we only used the cells from the 'Cortex' region. No cells were filtered from the **STARmap** dataset. Further, each dataset was normalized by dividing the counts within each cell by the total number of transcripts within that cell, scaling by the median number of transcripts per cell, and log1p transformed. Similar to the scRNA-seq data, we scaled the spatial data using the SciPy Python package[25].

It is important to note that in all experiments, the scaled datasets are used as input for the alignment part, while the prediction is applied using the normalized version of the scRNA-seq dataset (Equation 8.10).

8.2.5 CROSS VALIDATION

We evaluated the prediction performance of all methods using a leave-one-gene-out cross validation. For a set of N shared genes between the spatial and the scRNA-seq datasets, one gene is left out and the remaining $N-1$ genes are used for integration and prediction of the left-out gene. The prediction is then evaluated by comparing the measured and predicted spatial profiles of the left-out-gene.

For the **STARmap_AllenVISp** dataset pair, we applied a more challenging cross validation setup. Similar to the leave-one-gene-out setup, for a set of N shared genes, one gene is left out to be predicted. From the remaining $N-1$ genes we excluded the 100 genes that are most correlated (absolute Pearson correlation) with the left-out gene. The remaining $N-101$ genes are then used for the integration and prediction of the left-out genes.

8.2.6 BENCHMARKED METHODS

We compared the performance of SpaGE versus three state-of-the-art methods for data integration: Seurat, Liger, and gimVI. Seurat and Liger are available as R packages, while gimVI is available through the scVI Python package[26]. We were not able to include Harmony in the comparison, as the code to predict unmeasured gene expression is not available. During the benchmark, all methods were applied using their default settings, or the settings provided in the accompanying examples or vignettes. Data normalization and scaling were

performed using the built-in functions in each package, *NormalizeData* and *ScaleData* functions in Seurat, *normalize* and *scaleNotCenter* functions in Liger, while gimVI implicitly preprocess the data while computing.

8.2.7 MORAN'S I STATISTIC

The Moran's I statistic[27] is a measure of spatial autocorrelation calculated for each spatial gene. The Moran's I values can range from -1 to 1, where a value close to 1 indicates a clear spatial pattern, and a value close to 0 indicates random spatial expression, while a value close to -1 indicated a chess board like pattern. We calculated the Moran's I using the following equation:

$$I = \frac{N}{W}\frac{\sum_i \sum_j w_{ij}\,(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{8.12}$$

Where x is the gene expression array, $\bar{x}$ is the mean expression of gene x, N is the total number of spatial cells, w_{ij} is a matrix containing spatial weights with zeros on the diagonal, and W is the sum of w_{ij}. We calculated the spatial weights w_{ij} using the XY coordinates of the spatial cells, for each cell we calculated the kNN using the spatial coordinates (k=4). We assigned w_{ij} = 1 if j is in the nearest neighbors of i, otherwise w_{ij} = 0.

8.2.8 DOWN-SAMPLING

For the 994 shared genes in the **STARmap_AllenVISp** dataset pair, we first selected the top 50 spatial genes with high Moran's I statistic values to be used as test set. For the remaining 944 genes, we calculated the pairwise Pearson correlation using the scRNA-seq dataset. If the absolute value of the correlation of two genes is larger than 0.7, we removed the gene with the lower variance. After removing highly correlated genes, we sorted the remaining genes according to their expression variance in the scRNA-seq dataset. We selected the top 10, 30, 50, 100, 200 and 500 genes with high variance, these genes were used for alignment of the two datasets and prediction of the expression of the test genes. The prediction performance of these gene sets was compared with using all 944 genes.

We applied the same down-sampling criteria on the 9,751 shared genes in the **seqFISH_AllenVISp** dataset pair, except for two differences: (i) the 50 spatial genes used as test set were selected as the top predicted genes in the leave-one-gene-out cross validation experiment, (ii) after removing correlated genes, we selected sets of the top 10, 30, 50, 100, 200, 500, 1000, 2000, 5000 and 7000 most variable genes, as well as all 9,701 genes.

8.2.9 CELL-TYPE MARKER GENES

To evaluate the performance of SpaGE per cell type, we defined sets of marker genes for four major brain cell types: Inhibitory neurons, Excitatory neurons, Astrocytes and Oligodendrocytes. The marker genes of the **osmFISH** dataset were directly obtained from the original paper[5]. For the **STARmap** and **MERFISH** datasets, we used the *FindMarkers* function from the Seurat R package to define the top 20 differentially expressed genes per cell type, comparing one cell type vs the rest using a two-sided Wilcoxon rank sum test and the Bonferroni method for multiple test correction, with min.pct = 0.25 and logfc.threshold = 0.25.

8.2.10 A MODEL TO PREDICT TRUSTWORTHINESS OF THE SPAGE PREDICTION

To determine whether we can trust a predicted spatial pattern by SpaGE, we trained a logistic regression model that predicts the trustworthiness of the predicted signal from four characteristics of the data: (i) the Moran's I statistic of the predicted spatial gene expression (pMI_i), (ii) the mean μ_i and (iii) variance σ_i of the expression of that gene in the scRNA-seq data, and (iv) the percentage of cells expressing that gene in the scRNA-seq data (e_i). The trustworthiness, Y_i, used to train the model, is determined from the Spearman correlation between the SpaGE-predicted spatial pattern and the measured spatial pattern, i.e. correlations above the median correlation are considered to be trustworthy. This gives the following logistic regression model:

$$Y_i \sim pMI_i + \mu_i + \sigma_i + e_i \tag{8.13}$$

Note that the inputs to the model can be determined without the need to have access to the measured spatial expression of the gene, and consequently the model can be used to evaluate whether the predicted spatial pattern of expression of an unmeasured spatial gene is to be trusted or not.

8.3 RESULTS

8.3.1 SPAGE OVERVIEW

We developed SpaGE, a platform that enhances the spatial transcriptomics data by predicting the expression of unmeasured genes, from a dissociated scRNA-seq data from the same tissue (Figure 8.1). Based on the set of shared genes, we align both datasets using the domain adaptation method PRECISE[20], to account for technical differences as well as gene detection sensitivity differences. PRECISE geometrically aligns linear latent factors computed on each dataset and finds gene combinations expressed in both datasets. These gene combinations thus define a common latent space and can be used to jointly project both datasets. Next, in this common latent space, we use the kNN algorithm to define the neighborhood of each cell in the spatial data from the scRNA-seq cells. These neighboring scRNA-seq cells are then used to predict the expression of spatially unmeasured genes. Finally, we end up with the full gene expression profile of each cell in the spatial data.

The alignment step is the most crucial step in the pipeline of SpaGE. For this purpose, we use PRECISE, a domain adaptation method previously proposed to predict the drug response of human tumors based on pre-clinical model such as cell lines and mouse models. We adapted PRECISE to the task of integrating the spatial data with the scRNA-seq data by defining the common aligned subspace between both datasets (Figure 8.1). PRECISE takes as input the expression matrix of both datasets, having the same set of (overlapping) genes but measured differently and within different cells. As we are aiming to fit each spatial cell to the most similar scRNA-seq cells, we may refer to the spatial dataset as the 'query' and the scRNA-seq dataset as the 'reference'. First, PRECISE obtains a lower dimensional space for each dataset separately using a linear dimensionality reduction method, such as Principal Component Analysis (PCA). Next, the two independent sets of principal components (*PCs*) are aligned by applying a singular value decomposition. We align the two sets of principal components using the singular vectors to obtain the aligned components, named principal vectors (*PVs*). These *PVs* are sorted in decreasing order based on their similarity between the reference and the query datasets. This allows us to filter out dissimilar or noisy signals, by discarding *PVs* with relatively low similarity, thus keeping only the common latent space

(Methods). The principal vectors of the reference dataset (PV_r) are considered as the aligned latent space. We project both datasets on PV_r to obtain the new aligned versions used for the kNN prediction.

We performed SpaGE on six dataset pairs from different regions in the mouse brain, varying in the number of cells and the number of spatially measured genes, summarized in Table 8.1. To show the alignment performance, we calculated the cosine similarity between the *PCs* and the *PVs* i.e. before and after the alignment. Across all six dataset pairs, we observed that indeed the relation between the *PCs* is not one-to-one, as these *PCs* are obtained from two different datasets (Supplementary Figure 8.1 and 8.2). However, after alignment using PRECISE, the diagonal cosine similarity between the *PVs* is maximized showing a one-to-one relationship between the *PVs* of both datasets. Supplementary Figure 8.1A shows the diagonal cosine similarity before and after PRECISE (i.e. between *PCs* and *PVs*) across all dataset pairs, showing a relatively large increase in similarity after the alignment using PRECISE. As we used only the informative *PVs*, the final number of *PVs* varied across datasets (Supplementary Table 8.1) and, as a result, the amount of explained variance for each dataset varied, from ~6% for the **seqFISH+** dataset to ~94% for the **osmFISH** dataset.

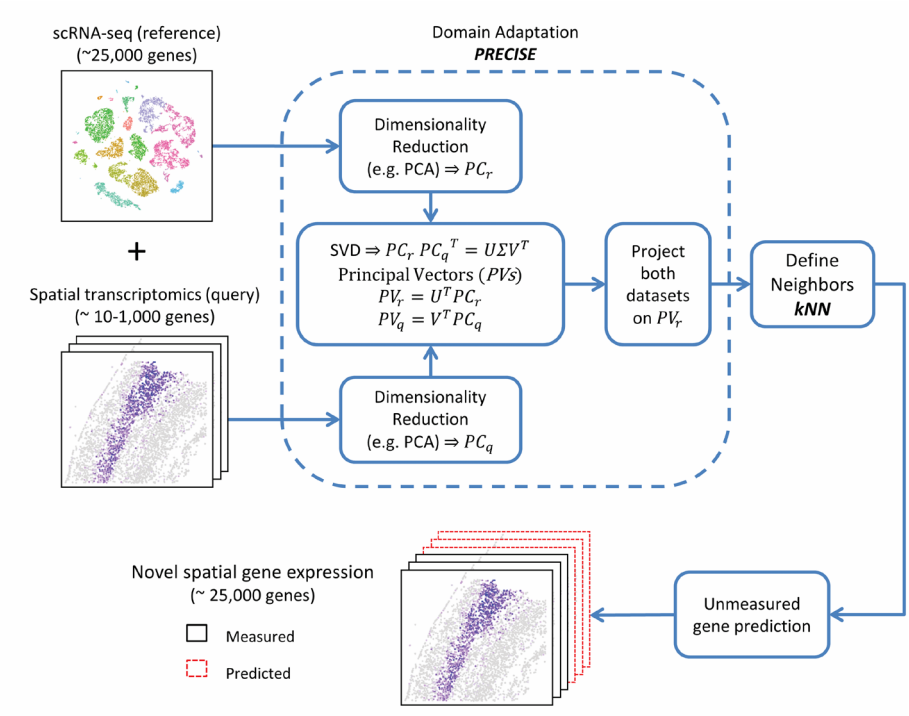


Figure 8.1 SpaGE pipeline. SpaGE takes as input two datasets, a scRNA-seq dataset and a spatial transcriptomics dataset measured from the same tissue. SpaGE uses gene combinations of equal significance in both datasets to predict spatial locations of unmeasured genes. Using PRECISE, SpaGE finds directions that are important for both datasets, by making use of a geometrical alignment of the independent PCs to produce the PVs. SpaGE aligns both datasets by projecting on the PVs of the reference dataset. Using the aligned datasets, SpaGE applies kNN prediction to define new gene expression patterns for spatially unmeasured genes, predicted from the dissociated scRNA-seq data. Each spatial cell can be enhanced by having the expression of the whole transcriptome.

Another interesting feature of SpaGE is the ability to interpret the most contributing genes defining the latent integration space. In general, these genes are highly variable and in most cases are related to cell type differences. A good example is the integration of the **osmFISH_Zeisel** dataset pair, in which the top six contributing genes are *Tmem2*, *Mrc1*, *Kcnip2*, *Foxj1*, *Apln* and *Syt6* (Methods). These genes are related to six different cell categories previously defined in the **osmFISH** paper[5]: Oligodendrocytes, Immune cells, Inhibitory neurons, Ventricle, Vasculature and Excitatory neurons, respectively.

We further illustrate the quality of the alignment by examining the overlap in the top contributing genes for the *PCs* (before PRECISE) and the *PVs* (after PRECISE). Using the **STARmap_AllenVISp** dataset pair, we obtained the top 50 contributing genes for the *PCs* of the **STARmap** data and the *PCs* of the **AllenVISp** data. These two sets shared only 2 genes out of 50. After alignment, the shared genes, between the top 50 contributing genes for the *PVs* of the **STARmap** data and the *PVs* of the **AllenVISp** data, increased to 12 genes. Also, we applied GO enrichment on these top contributing gene sets in each case using PANTHER (http://pantherdb.org/, Fisher exact test with Bonferroni multiple test correction). The **STARmap** *PCs* and the **AllenVISp** *PCs* had 9 enriched biological processes each, sharing 3 processes in common (Supplementary Table 8.2). While the **STARmap** *PVs* and the **AllenVISp** *PVs* had 27 and 41 enriched biological processes, respectively, sharing 12 processes in common. Interestingly many of them related are to regulation processes, such as regulation of biological process, cell population proliferation, metabolic processes, cell motility, locomotion, and cellular component movement.

8.3.2 SPAGE OUTPERFORMS STATE-OF-THE-ART METHODS ON THE STARMAP DATASET

Using the first dataset pair **STARmap_AllenVISp**, we applied SpaGE to integrate both datasets and predict unmeasured spatial gene expression patterns. In order to evaluate the prediction, we performed a leave-one-gene-out cross validation (Methods). The **STARmap_AllenVISp** dataset pair shares 994 genes. In each cross-validation fold, one gene is left out and the remaining 993 genes are used as input for SpaGE to predict the spatial expression pattern of the left-out gene. We evaluated the prediction performance by calculating the Spearman correlation between the original measured spatially distributed values and the predicted values of the left-out gene. We performed the same leave-one-gene-out cross validation using Seurat, Liger and gimVI, to benchmark the performance of SpaGE. Results show a significant improvement in performance for SpaGE compared to all three methods (p-value <0.05, two-sided paired Wilcoxon rank sum test), with a median Spearman correlation of 0.125 compared to 0.083, 0.067 and 0.035 for Seurat, Liger and gimVI, respectively (Figure 8.2A).

Further, we compared the Spearman correlation of SpaGE versus the state-of-the-art methods per gene, to obtain a detailed evaluation. Results show better performance of SpaGE across the majority of genes, but not all (Figure 8.2B-D). Next, we visually compared a few genes that had high correlations for each method. For the top three predicted genes of SpaGE (*Pcsk2*, *Pgm2l1* and *Egr1*), Seurat obtained a good prediction as well, as these three genes are in the top 10 predicted genes of Seurat. Liger failed to predict *Egr1*, while gimVI failed to predict *Pgm2l1* and *Egr1* (Supplementary Figure 8.3A). We further looked for examples where other methods obtained higher correlations than SpaGE, excluding the top 10 predicted genes by SpaGE. Compared to Seurat, SpaGE similarly predicted the expression of *Arpp19*, but predicted relatively higher contrast patterns for *Pcp4* and *Arc* (Supplementary Figure 8.3B). Compared to Liger, SpaGE similarly predicted the expression of *Mobp*, higher contrast pattern for *Hpcal4*, and better predicted the spatial pattern of *Tsnax* (Supplementary

Figure 8.3C). Compared to gimVI, SpaGE predicted a lower contrast pattern for *Arx*, a higher contrast pattern for *Snurf*, but failed to reproduce the measured spatial pattern for *Bcl6* (Supplementary Figure 8.3D). Remarkably, the predicted spatial patterns of SpaGE, for all three genes, are more in agreement with the data from the Allen Brain Atlas, suggesting that these genes were not accurately measured in the **STARmap** dataset.

Although the correlation values are in general low, SpaGE is capable of accurately reconstructing genes with clear spatial pattern in the brain. Figure 8.2E shows a set of genes known to have spatial patterns (previously reported by Seurat, Liger and gimVI). In this set of genes, Seurat and Liger are performing well, except that Liger produced a lower contrast expression pattern in some cases (e.g. *Lamp5* and *Bsg*). gimVI produced good prediction for *Lamp5*, however, gimVI was not able to predict the correct gene patterns for the other genes.

To obtain a better understanding and interpretation of these correlation values, we evaluated the effect of the kNN algorithm on the prediction performance. To do so, we divided the **AllenVISp** dataset into two stratified folds ensuring an equal composition of cell types. We used one-fold to predict genes in the other fold using the shared genes. Note that this does not require an alignment (PRECISE), so we can test the influence of the kNN regression. We applied a leave-one-gene-out cross validation using the same set of 994 shared genes of the **STARmap_AllenVISp** dataset pair, which resulted in a median Spearman correlation of 0.551 (Supplementary Figure 8.4A). While the performance is clearly better compared to that of SpaGE using the **STARmap_AllenVISp** dataset pair (median Spearman correlation = 0.125), it shows that the kNN regression is partially responsible for reduced correlation values.

To investigate the influence of the correlation metric, we tested also the Pearson and Kendall correlation measures, which showed that the highest correlation values are obtained when using the Spearman correlation (Supplementary Figure 8.4B). Next, we were interested how well SpaGE could predict when there was no difference between measurement modalities (here, spatial and scRNA-seq). Therefore, we used SpaGE to integrate the **Zeisel** and **AllenSSp** datasets, representing two scRNA-seq measured datasets from the same brain region. Using the leave-one-gene-out cross validation and the same shared genes of the **STARmap_AllenVISp** dataset pair, we obtained a median Spearman correlation of 0.303 (query: **Zeisel**, reference: **AllenSSp**) and 0.331 (query: **AllenSSp**, reference: **Zeisel**) (Supplementary Figure 8.4B). These correlation values suggest that the observed correlation values obtained when applying SpaGE on spatial and scRNA-seq datasets are not as low as they appear.

Additionally, although it is important to accurately predict the expression of all genes, genes with distinct spatial patterns are more important to accurately predict compared to non- or randomly expressed genes. To quantify the existence of spatial patterns, we calculate the Moran's I statistics of each gene using the original **STARmap** spatial data (Methods). We compared the prediction performance of each gene with the corresponding Moran's I value. For SpaGE, Seurat and Liger, we observed a positive relationship between the prediction performance and the Moran's I values, i.e. genes with spatial patterns are better predicted (Supplementary Figure 8.5A-C). On the other hand, gimVI performed worse on genes with high Moran's I statistics (Supplementary Figure 8.5D).

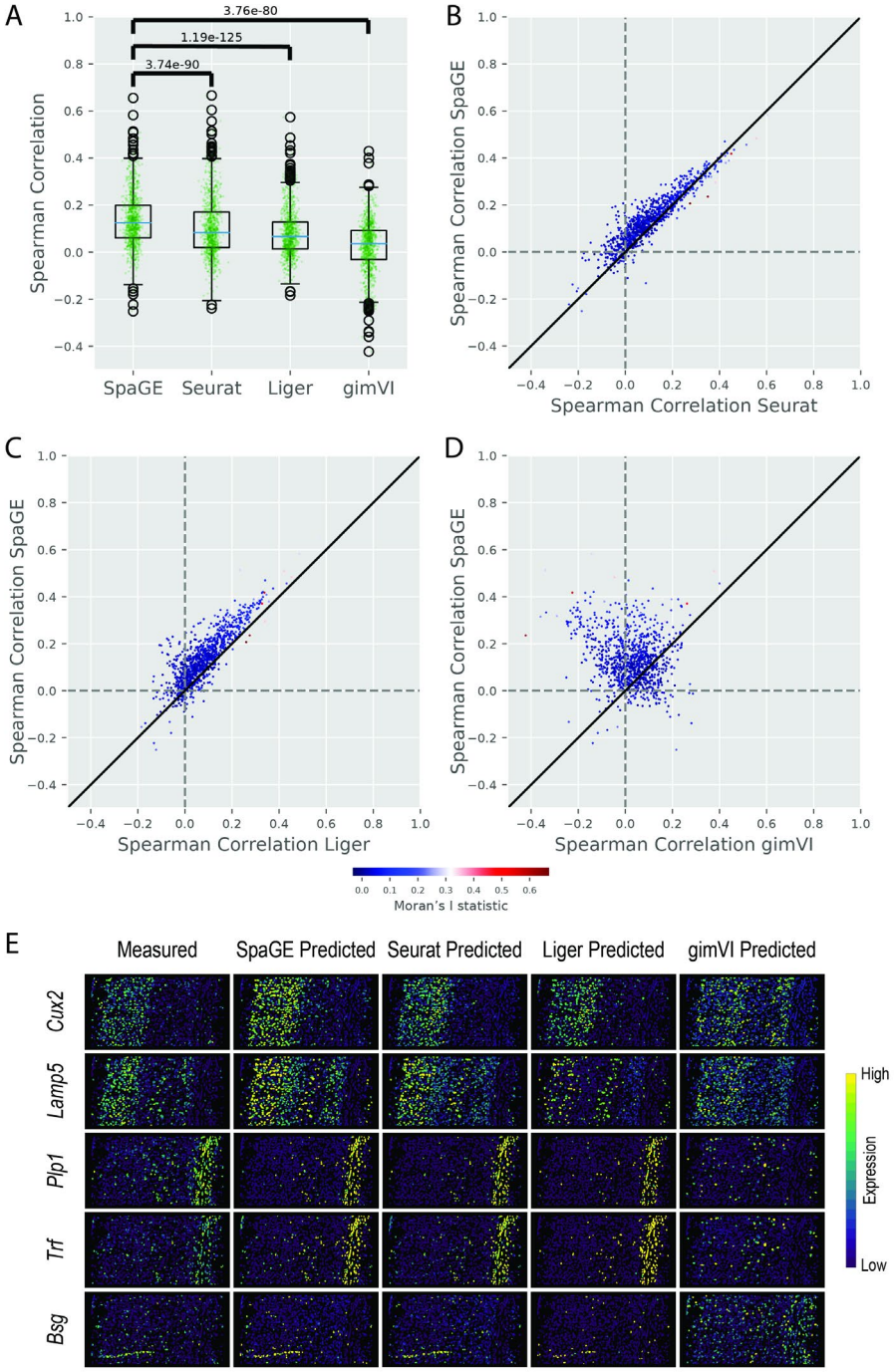


Figure 8.2 Prediction performance comparison for the STARmap_AllenVISp dataset pair. (**A**) Boxplots showing the Spearman correlations for the leave-one-gene-out cross validation experiment for each method. The blue lines show the median correlation across all genes with a better performance for SpaGE. The green dots show the correlation values for individual genes. The *P*-values show the significant difference between all correlation values of SpaGE and each other method, using a paired Wilcoxon rank-

sum test. (**B–D**) Detailed performance comparison between SpaGE and (**B**) Seurat, (**C**) Liger, (**D**) gimVI. These scatter plots show the correlation value of each gene across two methods. The solid black line is the $y = x$ line, the dashed lines show the zero correlation. Points are colored according to the Moran's I statistic of each gene. All scatter plots show that the majority of the genes are skewed above the $y = x$ line, showing an overall better performance of SpaGE over other methods. (**E**) Predicted expression of known spatially patterned genes in the **STARmap** dataset. Each row corresponds to a single gene having a clear spatial pattern. First column from the left shows the measured spatial gene expression in the **STARmap** dataset, while other columns show the corresponding predicted expression pattern by SpaGE, Seurat, Liger and gimVI, using the leave-one-gene-out cross validation experiment. Prediction is performed using the **AllenVISp** dataset.

Further, we evaluated the prediction performance of all methods using a more challenging cross validation setup. Compared to the (traditional) leave-one-gene-out setup, the left-out gene is predicted using less shared genes in this set up, i.e. we removed the (100) most correlated genes with the left-out gene from the training set (Methods). This more challenging evaluation did result in comparable prediction performance to the leave-one-gene-out setup, with roughly the same differences and ranking across all methods (Supplementary Figure 8.6A). In addition, we evaluated how well a gene can be predicted when using less shared genes in general. First, we selected a fixed test set of 50 genes, next we down-sampled the remaining set of 944 shared genes in a guided manner (Methods). For down-sampled shared genes sets of 10, 30, 50, 100, 200, 500 and all 944 genes, SpaGE performance always increases with the number of shared genes as expected (Supplementary Figure 8.6B).

8.3.3 SPAGE PREDICTS UNMEASURED SPATIAL GENE PATTERNS THAT ARE INDEPENDENTLY VALIDATED

After validating SpaGE to accurately predict the spatially measured genes, we applied SpaGE to predict new unmeasured genes for the spatial data, with the aim to define novel spatial gene patterns. We illustrate SpaGE's capability of such task using the **STARmap_AllenVISp** dataset pair. First, during the leave-one-gene-out cross validation, SpaGE was able to produce the correct spatial pattern for *Rorb*, *Syt6* and *Tbr1* (Figure 8.3). These three genes were originally under-expressed, possibly due technical noise or low gene detection sensitivity in the **STARmap** dataset. Our predictions using SpaGE are in agreement with the highly sensitive cyclic smFISH dataset (**osmFISH**[5]) measured from the mouse somatosensory cortex, a similar brain region in terms of layering structure to the visual cortex measured by the **STARmap** dataset. Further, using SpaGE, we were able to obtain novel spatial gene patterns for five genes not originally measured by the **STARmap** dataset, showing clear patterns through the cortical layers (Figure 8.4). These predicted patterns are supported by the Allen Brain Atlas in-situ hybridization (ISH).

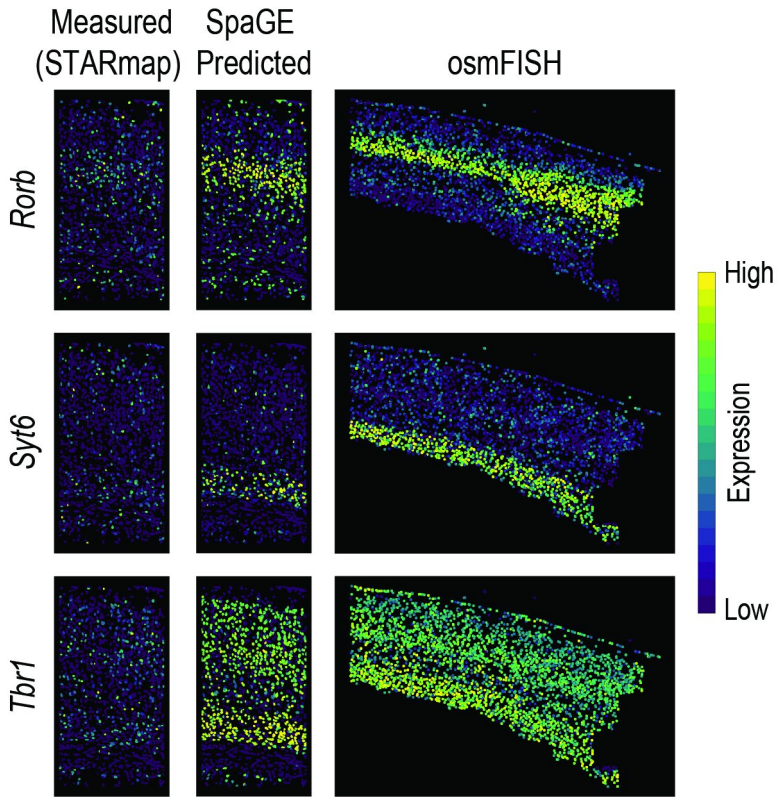


Figure 8.3 SpaGE accurately predicted the expression of *Rorb*, *Syt6* and *Tbr1* in agreement with the osmFISH data. These three genes (shown in rows) were wrongly measured in the original **STARmap** data (shown in the left column). Using the **STARmap_AllenVISp** dataset pair, SpaGE was able to reconstruct the correct spatial gene expression patterns (middle column). These predicted patterns are in agreement with the measured gene expression patterns measure by the **osmFISH** dataset (right column), a highly sensitive single-molecule technology.

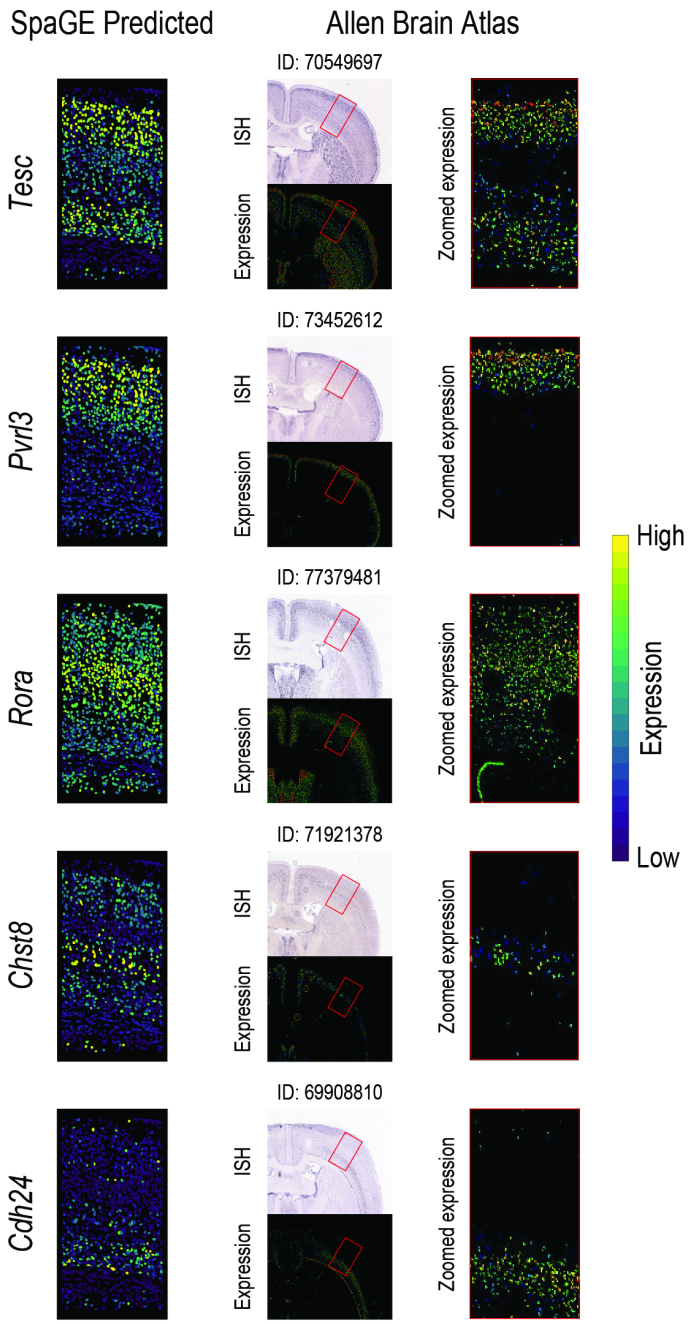


Figure 8.4 Novel gene expression patterns for five genes not originally measured by the STARmap dataset, validated using the Allen Brain Atlas *in-situ* hybridization ISH. The left column shows the predicted spatial patterns using SpaGE for these five genes (shown in rows). The middle column shows the Allen Brain Atlas ISH data for each gene, stating the image ID on top of each tissue section. The red rectangle highlights the corresponding brain region measured by the **STARmap** dataset. The right column shows a zoomed-in view of the region highlighted using this red rectangle, showing an agreement with the expression patterns predicted by SpaGE.

To quantitatively evaluate the predicted spatial patterns for non-measured genes, we trained a logistic regression model to estimate whether a predicted spatial gene expression can be trusted or not (Methods). We used 3 statistical features from the scRNA-seq data, in addition to the Moran's I statistic of the *predicted* spatial pattern. When training this model, we used the Spearman correlation between the SpaGE-predicted spatial pattern and the measured spatial pattern to determine whether a gene can be trusted or not, i.e. we assumed that correlations above the median correlation are trustworthy. Using the 994 shared genes of the **STARmap_AllenVISp** dataset pair, we obtained an average accuracy of 0.71 for a stratified 2-folds cross validation. Next, we trained the model using all genes and applied it to the estimated gene patterns in Figures 8.3 and 8.4. This model judged the predicted patterns of *Rorb*, *Tbr1*, *Tesc*, *Pvlr3* and *Rora*, trustworthy, and the patterns for *Syt6*, *Chst8* and *Cdh24* were not. Interestingly, when inspecting the model's coefficients we found that the Moran's I statistic of the predicted spatial pattern had the largest contribution.

8.3.4 SpaGE Predictions Improve with Deeply Sequenced Reference Dataset

We wanted to test the effect of changing the reference scRNA-seq data on the spatial gene expression prediction. Here, we used the **osmFISH** dataset which represents a different challenge compared to the **STARmap** dataset. On one hand, the **osmFISH** dataset has a relatively higher gene detection sensitivity, but on the other hand, the **osmFISH** dataset includes only 33 genes. First, we evaluated the **osmFISH_Zeisel** dataset pair, in which we integrated the **osmFISH** dataset with a reference scRNA-seq dataset from the same lab[24]. We performed leave-one-gene-out cross validation similar to the **STARmap** dataset. Compared to other methods, SpaGE has significantly better performance (p-value <0.05, two-sided paired Wilcoxon rank sum test), with a median Spearman correlation of 0.203 compared to 0.007, 0.090 and 0.133 for Seurat, Liger and gimVI, respectively (Figure 8.5A, Supplementary Figure 8.7A). For a more detailed comparison per gene: SpaGE is performing better on the majority of genes compared to Liger and gimVI, while compared to Seurat, SpaGE has better performance across all genes (Supplementary Figure 8.7B-D). We further investigated the relation between the prediction performance and the Moran's I statistics of the originally measured genes. Similar to the **STARmap** data, for SpaGE and Seurat, we found a positive relationship, i.e. the performance is higher for genes with distinct spatial patterns. However, Liger and gimVI have a negative relationship (Supplementary Figure 8.8).

Next, we tested the performance of all methods using the **AllenVISp** dataset as reference for the **osmFISH** dataset, similar to the **STARmap** dataset. For the **osmFISH_AllenVISp** dataset pair, we observed similar conclusions where SpaGE has significantly better performance compared to other methods, with a median Spearman correlation of 0.203 compared to 0.014, 0.082 and 0.162 for Seurat, Liger and gimVI, respectively (Figure 8.5A, Supplementary Figure 8.9A). SpaGE has better performance across all genes compared to Seurat and Liger, while gimVI is performing better on a few genes (Supplementary Figure 8.9B-D). All four methods have a positive relationship between their prediction performance and the Moran's I statistics of the measured genes (Supplementary Figure 8.10). These results show how the reference dataset can affect the prediction. Compared to the **Zeisel** dataset, the **AllenVISp** is more deeply sequenced data, with the average number of detected transcripts per cell being ~140x more than the **Zeisel** dataset (Supplementary Figure 8.11A-B). However, not all methods benefit from this, as for Seurat and Liger, the prediction performance using the **AllenVISp** or the **Zeisel** datasets is quite similar (Figure 8.5A). On the other hand, SpaGE and gimVI get an increase in performance across all genes, although the median correlation for SpaGE remains the same. Similar to the **STARmap** dataset, we tested the performance of the KNN regression within the **AllenVISp** dataset only (excluding

the alignment procedure), when using only the 33 genes of the **osmFISH** dataset. In this case, we obtained a median correlation of 0.289 (Supplementary Figure 8.4A), when predicting the expression of genes in the scRNA-seq data from one-fold to the other, which is slightly higher than SpaGE (0.203) predicting **osmFISH** patterns. This result shows that the alignment of the spatial and scRNA-seq data using SpaGE is performing well, as the overall performance is comparable with predictions within the same dataset.

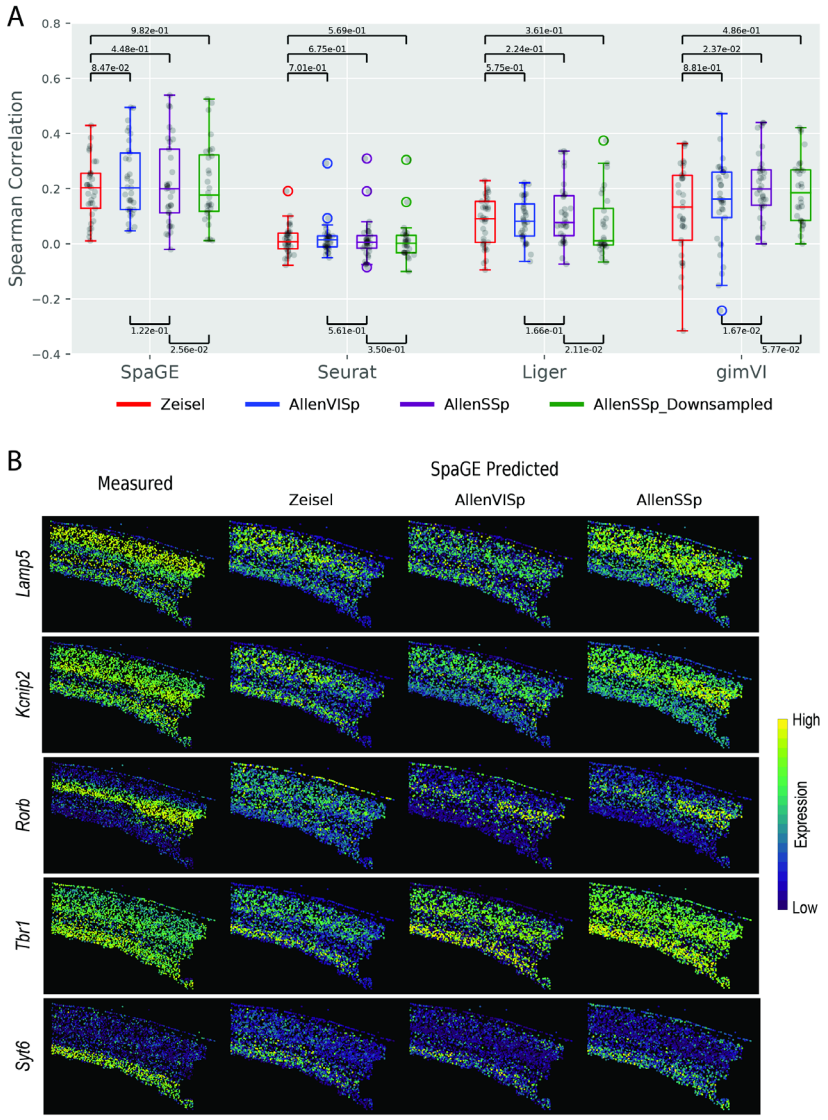


Figure 8.5 Prediction performance comparison for the osmFISH dataset using different reference scRNA-seq datasets. (**A**) Boxplots showing the Spearman correlations for the leave-one-gene-out cross validation experiment for each method using four different scRNA-seq datasets, **Zeisel**, **AllenVISp**, **AllenSSp** and **AllenSSp_Downsampled**. The median correlations shows a better performance for SpaGE in all dataset pairs. The black dots show the correlation values for individual genes. The *P*-values are obtained using a paired Wilcoxon rank-sum test. SpaGE showed a performance

improvement when using the **AllenVISp** over the **Zeisel** data. Although the median correlation is the same, the overall correlation range did improve. Also, gimVI clearly benefits from using the **AllenVISp** and the **AllenSSp** datasets over the **Zeisel** dataset. All methods have decreased performance when using the **AllenSSp_Downsampled** data compared to the original **AllenSSp** data. (**B**) Predicted expression of known spatially patterned genes in the **osmFISH** dataset using different reference scRNA-seq datasets. Each row corresponds to a single gene having a clear spatial pattern. First column from the left shows the measured spatial gene expression in the **osmFISH** dataset, while the second, third and fourth columns show the corresponding predicted expression pattern by SpaGE using **Zeisel**, **AllenVISp** and **AllenSSp** datasets, respectively. Changing from **Zeisel** to **AllenVISp** (deeply sequenced data) improved the prediction, while matching the brain region using the **AllenSSp** improved the prediction further.

While the **AllenVISp** is a deeply sequenced reference dataset, it has been measured from a different brain region than the **osmFISH** dataset (Table 8.1). Therefore, we decided to use a third reference dataset, **AllenSSp**, which has roughly the same sequencing depth as the **AllenVISp** (Supplementary Figure 8.11B-C) but is measured from the somatosensory cortex, similar to the **osmFISH** dataset. We evaluated the prediction performance of all four tools for the new dataset pair **osmFISH_AllenSSp**. SpaGE obtained a better performance with a median Spearman correlation of 0.199 compared to 0.006 and 0.077 for Seurat and Liger, respectively, while gimVI has similar performance to SpaGE with a median Spearman correlation of 0.199 (Figure 8.5A, Supplementary Figure 8.12A). SpaGE has a better performance across almost all genes compared to Seurat and Liger, while gimVI performed better than SpaGE for nearly half the genes (Supplementary Figure 8.12B-D). SpaGE, Liger and gimVI have positive relationship between the prediction performance and Moran's I statistics. However, Seurat has a negative relationship (Supplementary Figure 8.13).

Several sources of variation do exist between the **Allen** datasets and the **Zeisel** dataset; besides the sequencing depth, these datasets are, for example, generated in different labs and using different sequencing protocols. To separately test the effect of the sequencing depth of the reference scRNA-seq data on the prediction performance, we downsampled the **AllenSSp** dataset to a comparable number of transcripts per cell as the **Zeisel** dataset, using the scuttle R package. Compared to the original **AllenSSp** dataset, we obtained lower prediction performance across all methods when using the downsampled dataset (Figure 8.5A), clearly showing that a deeply sequenced reference dataset produces a better prediction. Interestingly, compared to the **Zeisel** dataset, the median performance using the downsampled **AllenSSp** dataset was lower for SpaGE, Seurat and Liger, but higher for gimVI.

Changing the brain region did not affect the overall performance of SpaGE (Figure 8.5A), however, the prediction of genes with known patterns did improve (Figure 8.5B). When we visually inspect these genes, we can clearly observe that the predicted spatial pattern improved when the sequencing depth of the reference set improves or becomes from a similar tissue. *Rorb* and *Tbr1* are clear examples, where the prediction using **Zeisel** was almost missing the correct pattern, this became clearer using the **AllenVISp** having a greater sequencing depth. Changing to a matching tissue adds further improves the predicted patterns of these genes (**AllenSSp**). Eventually, all five genes (*Lamp5*, *Kcnip2*, *Rorb*, *Tbr1* and *Syt6*) are more accurately predicted using the **AllenSSp** dataset. Moreover, we used the **AllenSSp** reference dataset to predict the spatial expression of 10 genes not originally measured by the **osmFISH** dataset, with clear patterns through the cortical layers (Figure 8.6). These predicted patterns are in agreement with the Allen Brain Atlas in-situ hybridization (ISH).

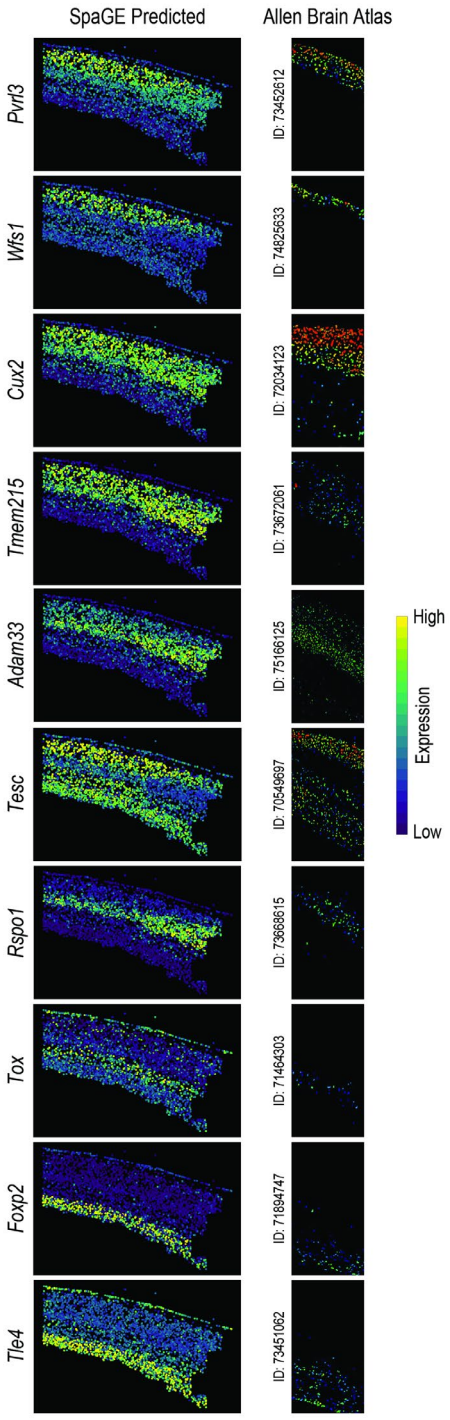


Figure 8.6 Novel gene expression patterns for 10 genes not originally measured by the osmFISH dataset, validated using the Allen Brain Atlas *in-situ* hybridization ISH. The left column shows the

predicted spatial patterns using SpaGE for these 10 genes (shown in rows). The right column shows the Allen Brain Atlas ISH expression for each gene, stating the image ID next to the tissue section, showing an agreement with the expression patterns predicted by SpaGE. These genes show clear expression to specific cortical layers (*Pvrl3* and *Wfs1*: layer 2/3; *Cux2*, *Tmem215* and *Adam33*: layer 2/3 and layer 4; *Rspo1*: layer 4; *Tesc*: layer 2/3 and layer 6; *Tox*: layer 5; *Foxp2* and *Tle4*: layer 6).

8.3.5 SPAGE IS SCALABLE TO LARGE SPATIAL DATASETS

So far, SpaGE showed good prediction performance in the leave-one-gene-out predictions, and was also able to predict correct spatial patterns of unmeasured genes within the spatial transcriptomic datasets. All these results were, however, obtained using a relatively small spatial datasets including only a few thousand cells (**STARmap** and **osmFISH**). This opens the question of how does SpaGE scale to large spatial datasets, comparable to the datasets measured nowadays. To assess the scalability of SpaGE, we used a large **MERFISH** dataset with >60,000 cells measured from the mouse brain pre-optic region, and integrated it with the corresponding scRNA-seq dataset published in the same study by **Moffit** et al[4]. The **MERFISH_Moffit** dataset pair shares 153 genes on which we applied the same leave-one-gene-out cross validation using all four methods. Similar to the previous results, SpaGE significantly outperformed all other methods (p-value <0.05, two-sided paired Wilcoxon rank sum test) with a median Spearman correlation of 0.275 compared to 0.258, 0.027 and 0.140 for Seurat, Liger and gimVI, respectively (Figure 8.7A). Per gene comparisons shows a clear advantage of SpaGE versus Liger and gimVI, but more comparable performance with Seurat (Figure 8.7B-D). The reported p-values are quite significant, however, it is important to note that the p-values are inflated due to the large sample size, which is also the case for the **STARmap** dataset.

Next to the overall performance across all genes, we evaluated the performance of SpaGE to predict marker genes of four major brain cell types: inhibitory neurons, excitatory neurons, astrocytes and oligodendrocytes (Methods). We observed that SpaGE had higher prediction performance for cell type marker genes compared to the overall performance across all genes (Figure 8.7E). Similar conclusion can be observed for the **STARmap** dataset (Supplementary Figure 8.14A), however, this is not the case for the **osmFISH** dataset because almost all 33 genes were cell type marker genes (Supplementary Figure 8.14B). Additionally, the ranking of the prediction performance across cell types is related to the cell type proportions observed in the data. For instance, the **MERFISH** dataset has approximately 38% inhibitory neurons, 18% excitatory neurons, 15% oligodendrocytes and 13% astrocytes, for which the median correlation per cell type is 0.587, 0.551, 0.402 and 0.398, respectively (Figure 8.7E). Compared to the pre-optic region, the cortex contains more excitatory neurons than inhibitory. This is directly reflected in the prediction performance of inhibitory and excitatory marker genes, where the latter have higher performance for the cortical datasets **STARmap** and **osmFISH** (Supplementary Figure 8.14).

Further, we compared the computation times of all four methods across all five dataset pairs. All experiments were run on a Linux HPC server but limited to a single CPU core, with 256 GB of memory, to be able to compare runtimes. For all methods, the calculated computation time includes the integration and the prediction time. Overall SpaGE has the lowest average computation time per gene, across all five dataset pairs (Figure 8.7F). For the large **MERFISH** dataset, SpaGE has a clear advantage compared to the other methods as the average computation time of SpaGE is ~30x, 63x and 45x faster than Seurat, Liger and gimVI, respectively. In terms of memory, SpaGE has the lowest memory usage across all five dataset pairs, while Seurat and Liger consumed memory the most (Figure 8.7F). Combined, these results show an overall advantage of SpaGE over other methods for larger datasets with higher prediction performance, lower computation time and less memory requirement.

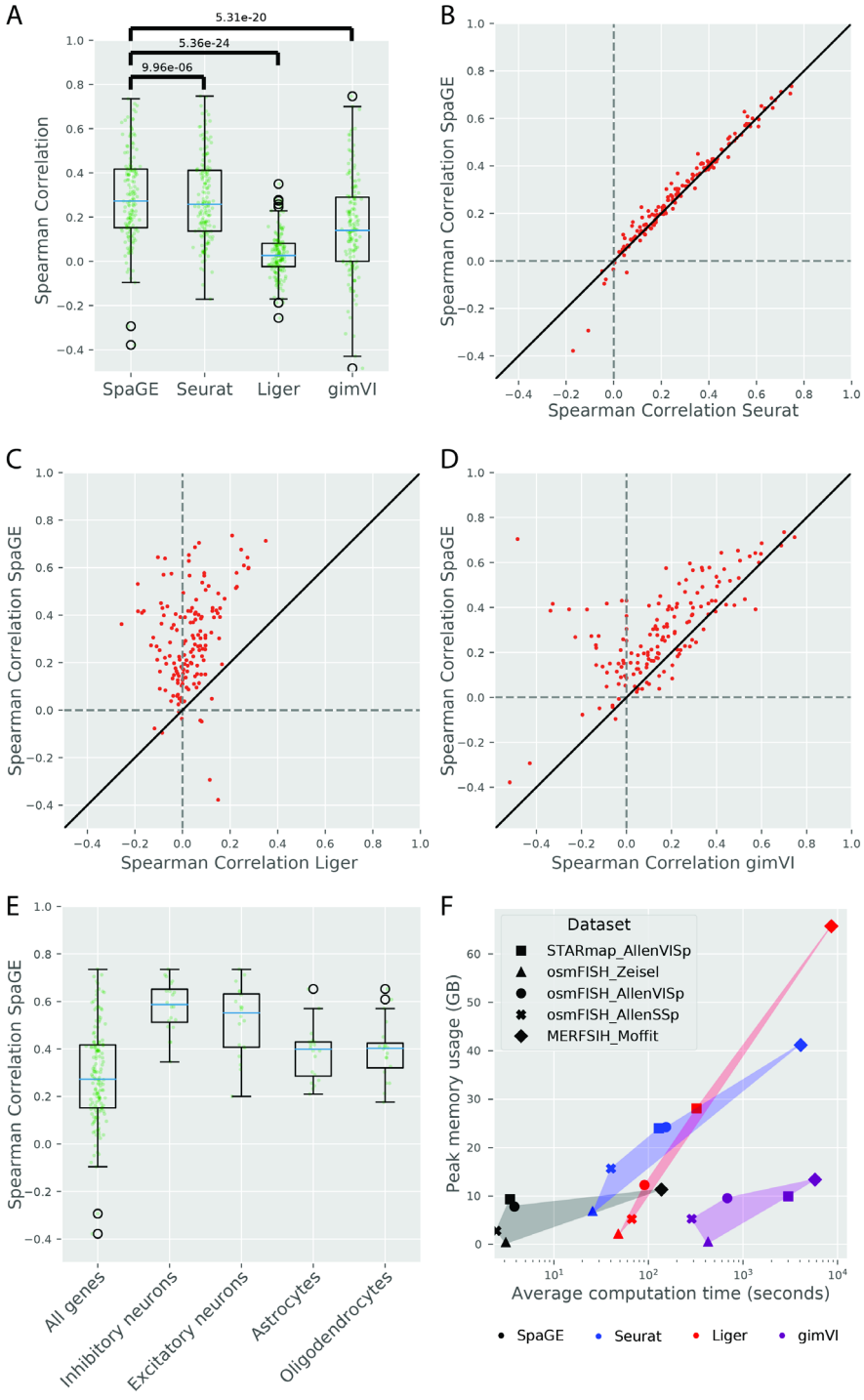


Figure 8.7 Prediction performance comparison for the MERFISH_Moffit dataset pair. (**A**) Boxplots showing the Spearman correlations for the leave-one-gene-out cross validation experiment for each method. The blue lines show the median correlation across all genes with a better performance for SpaGE.

The green dots show the correlation values for individual genes. The *P*-values show the significant difference between all correlation values of SpaGE and each other method, using a paired Wilcoxon rank-sum test. (**B–D**) Detailed performance comparison between SpaGE and (**B**) Seurat, (**C**) Liger, (**D**) gimVI. These scatter plots show the correlation value of each gene across two methods. The solid black line is the $y = x$ line, the dashed lines show the zero correlation. All scatter plots show that the majority of the genes are skewed above the $y = x$ line, showing an overall better performance of SpaGE over other methods. (**E**) Boxplots showing the prediction performance of SpaGE for cell type marker genes compared to the overall performance across all genes. (**F**) scatter plot showing the average computation time (log-scaled) per gene versus the peak memory usage. Methods are represented with different colors and dataset pairs are represented with different symbols. Points of the same method are highlighted for clarity.

8.3.6 INCREASING THE NUMBER OF SHARED GENES DOES NOT ALWAYS IMPROVE THE PREDICTION

To investigate whether the performance improves when having many more spatially measured genes, we tested SpaGE when applying it to the **seqFISH+** spatial dataset that measures up to 10,000 genes simultaneously. Using the **seqFISH_AllenVISp** dataset pair, we applied SpaGE using the leave-one-gene-out cross validation setup to predict the spatial expression of 9,751 shared genes. SpaGE produced a median Spearman correlation of 0.154, a minimum correlation of -0.170 and a maximum correlation of 0.716. This result is comparable to the other tested dataset pairs, showing robust performance of SpaGE.

However, with ~10,000 spatial genes, we expected a better performance as there are many more shared genes with which matching cells can be found in the scRNA-seq data. To further substantiate this, we compared the prediction performance of 494 overlapping genes between the **seqFISH_AllenVISp** and the **STARmap_AllenVISp** dataset pairs, both having the same scRNA-seq reference data. The performance when using the **seqFISH+** data, having ~10x more shared genes, was significantly higher than when using the **STARmap** data (p-value <0.05, two-sided paired Wilcoxon rank sum test) (Figure 8.8A). A detailed comparison per gene shows that the majority of the genes are indeed better predicted in the **seqFISH+** dataset (Figure 8.8B). However, when comparing the 21 overlapping genes between the **seqFISH_AllenVISp** and the **osmFISH_AllenVISp** dataset pairs, we obtained a contradicting result. The performance when using the **osmFISH** data (only 33 shared genes) was higher than when we used the **seqFISH+** data, for almost all 21 genes for which we could make this comparison (Figure 8.8C-D).

This opens the question whether having more measured spatial genes (and thus shared genes) is always beneficial to predict the spatial patterns of non-measured genes. To answer that, we performed a downsampling experiment similar to what we did with the **STARmap** data (Methods). We fixed 50 genes as test set and downsampled the remaining genes to sets of the top 10, 30, 50, 100, 200, 500, 1000, 2000, 5000, 7000 and 9,701 (all) highly varying genes as shared genes. The best prediction performance of SpaGE was obtained using 5000 genes, after which the performance decreased (Figure 8.8E). Apparently, having more genes includes more and more lowly varying, and thus noisy, genes into the matching process, which turns out to confuse the matching process and consequently lower the prediction performance.

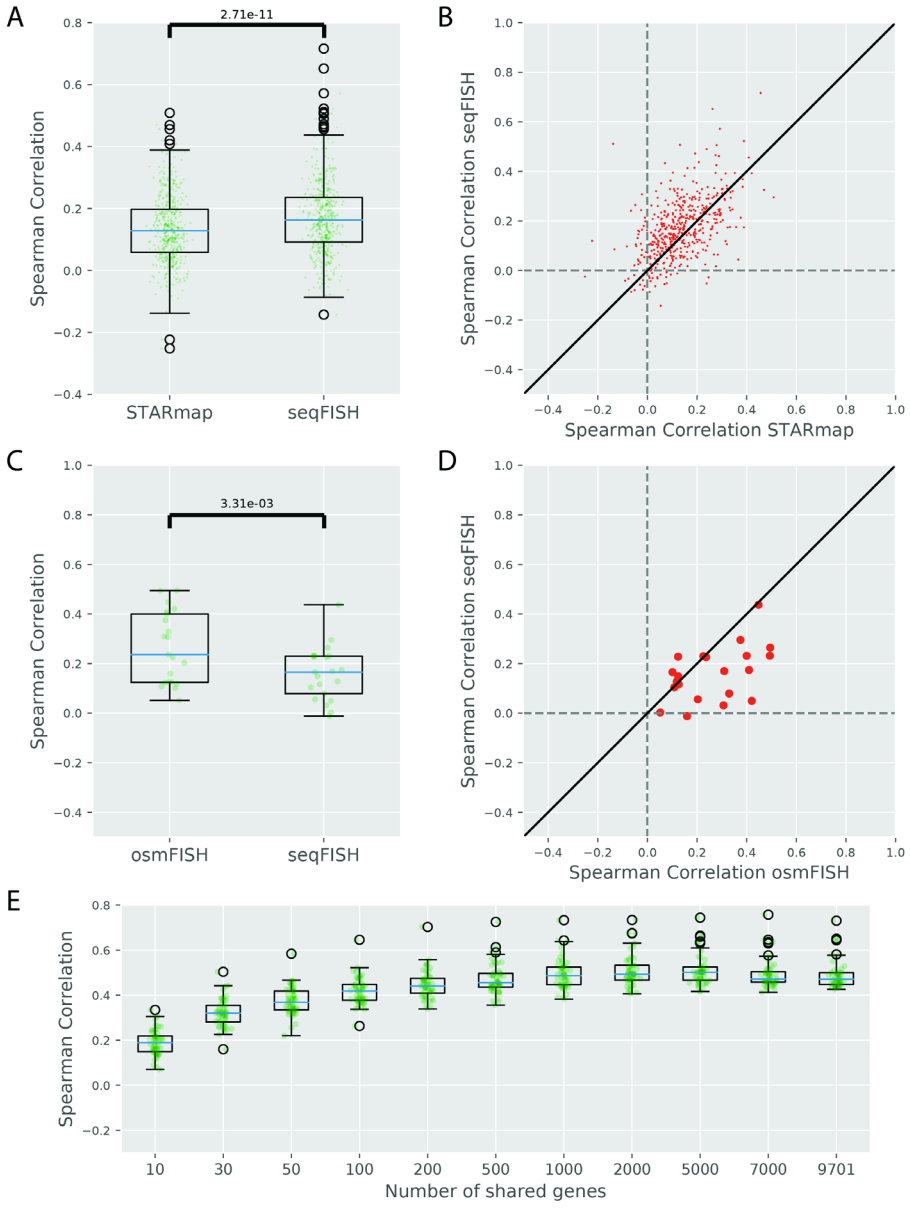


Figure 8.8 Prediction performance of SpaGE for the seqFISH_AllenVISp dataset pair. (**A,C**) Boxplots comparing the prediction performance of SpaGE for the shared genes between the **seqFISH** and the (**A**) **STARmap**, (**C**) **osmFISH** datasets, using the same **AllenVISp** dataset as reference during prediction. The blue lines show the median correlation across all genes. The green dots show the correlation values for individual genes. The *P*-value is obtained using a paired Wilcoxon rank-sum test. (**B, D**) Detailed performance comparison between **seqFISH** and (**B**) **STARmap**, (**D**) **osmFISH**. These scatter plots show the correlation value of each gene across two datasets. The solid black line is the $y = x$ line, the dashed lines show the zero correlation. (**E**) Boxplots showing the prediction performance of a test set of 50 genes, in terms of Spearman Rank correlations, using downsampled sets of 10, 30, 50, 100, 200, 500, 1000, 2000, 5000 and 7000 shared genes compared to using all 9,701 genes in the **seqFISH_AllenVISp** dataset pair.

8.4 DISCUSSION

We demonstrated the ability of SpaGE to enhance spatial transcriptomics data by predicting the expression of unmeasured genes based on scRNA-seq data collected from the same tissue. The ability of SpaGE to produce accurate gene expression prediction highly depends on the alignment part performed using PRECISE, which rotates the principal components of each dataset to produce principal vectors with high one-to-one similarity. Projecting the datasets to the latent space spanned by these principal vectors produces a proper alignment, making a simple kNN prediction sufficient to achieve accurate gene expression estimation.

During the alignment, SpaGE ignores principal vectors with low similarity which excludes uncommon and/or noisy signals. Despite the clear differences in the amount of explained variance for each dataset pair by the final set of principal vectors, SpaGE was able to capture the common sources of variation and produce good predictions of the spatial gene expressions across all dataset pairs. SpaGE captured ~6% of the variance for the **seqFISH+** dataset that measures ~10,000 genes spatially, but the majority of which are lowly variable in the mouse cortex, thus not contain enough information to contribute to the integration. On the contrary, SpaGE captured ~94% of the variance for the **osmFISH** dataset, which contains 33 known marker genes for various cell types in the mouse somatosensory cortex. Almost all these genes are highly variable and contain useful information for the integration.

We benchmarked SpaGE against three state-of-the-art methods for multi-omics data integration, using five different dataset pairs. These dataset pairs represent different challenges to the integration and prediction task, as they differ in gene detection sensitivity level and the number of spatially measured genes, which are the basis for the alignment. Increasing the number of shared genes should, in principle, eases the integration task and produces better predictions of spatial patterns of unmeasured genes. However, this is not always the case, as shown by the **seqFISH+** data, where adding more genes eventually also adds genes that have a relatively low variance, and thus are more probably noisy genes. This turns out to negatively influence the matching process and consequently decrease the prediction performance. Apparently, there is an optimum on the number of genes that need to be spatially measured when we want to predict spatial patterns of unmeasured genes. On the other hand, when measuring the spatial patterns measured for ~10,000 genes, it might not be necessary to predict spatial patterns of unmeasured genes as the initially spatially measured genes already cover most of the transcriptome of interest. Further, imaging-based spatial transcriptomic methods, with high gene detection sensitivity, may also improve the integration and prediction, as they are able to capture the majority of the genes even the ones with relatively low expression. On the other hand, integrating this high sensitivity data with scRNA-seq, which has lower sensitivity, can be more challenging. That is because the differences in gene expression are higher compared to integrating a sequencing-based spatial data with scRNA-seq data, both having comparable sensitivity.

Across all tested dataset pairs, SpaGE outperformed all methods producing better predictions for the majority of the genes. However, for few genes, SpaGE had lower prediction performance than other methods. Seurat produced good gene predictions for the **STARmap** and the **MERFISH** datasets, with similar predictions to SpaGE. However, Seurat had overall the lowest performance for the **osmFISH** dataset, with correlation close to 0, which shows that the performance of Seurat heavily decreased when there are very few shared genes, such as in the **osmFISH** dataset (33 genes). This problem is even more pronounced for Liger, as it performed relatively well for the **STARmap** dataset producing

good gene predictions, but has a decreased performance for both the **osmFISH** (33 genes) and the **MERFISH** (155 genes) datasets. On the other hand, gimVI performed relatively well for the **osmFISH** and the **MERFISH** datasets. However, gimVI had overall the lowest performance for the **STARmap** dataset, with inaccurate predictions for genes with spatial patterns such as *Cux2* and *Plp1*. This suggests that gimVI works well with imaging-based technologies having high gene detection sensitivity, but not with the sequencing-based technologies.

Next to the overall best performance, SpaGE is an interpretable algorithm as it allows to find the genes driving the correspondence between the datasets. The principal vectors, used to align the datasets to a latent space, show the contribution of each gene in defining this new latent space. Further, SpaGE is scalable to large spatial data with significantly lower computation time and memory requirement compared to the other methods, as shown on the **MERFISH** dataset having more than 60,000 cells measured spatially. Moreover, SpaGE is a flexible pipeline. Here we used PCA as the initial independent dimensionality reduction algorithm. However, this step can be replaced by any linear dimensionality reduction method.

SpaGE showed high prediction performance for cell type marker genes compared to the overall performance across all genes. These marker genes are often highly variable genes with clear spatial expression patterns. For example, *Cux2* and *Lamp5* represent two excitatory neurons marker genes with clear spatial patterns in the mouse cortex, which were well predicted by SpaGE. We also showed that the cell type proportions directly affect the prediction of the corresponding marker genes. However, the prediction of a marker gene is, in the first place, directly related to the existence of the corresponding cell type across both spatial and scRNA-seq datasets. For example, it is not possible to correctly predict the spatial expression of an astrocyte marker gene, if one or both datasets do not contain any astrocytes. In other words, it is better to measure both spatial and scRNA-seq datasets from the same sample, as we have seen in the **MERFISH_Moffit** dataset pair. However, datasets emerging from different samples but from matching tissue can still produce good spatial gene expression predictions if their cell type compositions are preserved.

We used the Spearman Rank correlation to quantitatively evaluate the predicted gene expressions. The overall evaluation showed relatively low correlations across all methods and all dataset pairs. These low correlations express the difficulty of the problem, as the predicted gene expressions are obtained from a different type of data. Given the low observed correlations, we developed a model that expresses whether we can trust a SpaGE-predicted spatial expression or not, which helps a user of SpaGE to interpret the correlations, improving the practicality of SpaGE. However, the Spearman correlation is not the optimal evaluation metric, as it does not always reflect the spatially predicted patterns, i.e. visual inspection showed good predictions for genes with known spatial pattern in the mouse cortex, while the correlation values were less than 0.2.

8.5 CONCLUSION

SpaGE presents a robust, scalable, interpretable and flexible method for predicting spatial gene expression patterns. SpaGE uses domain adaptation to align the spatial transcriptomics and the scRNA-seq datasets to a common space, in which unmeasured spatial gene expressions can be predicted. SpaGE is less complex and much faster when compared to other approaches and generalizes better across datasets and technologies.

8.6 DATA AVAILABILITY

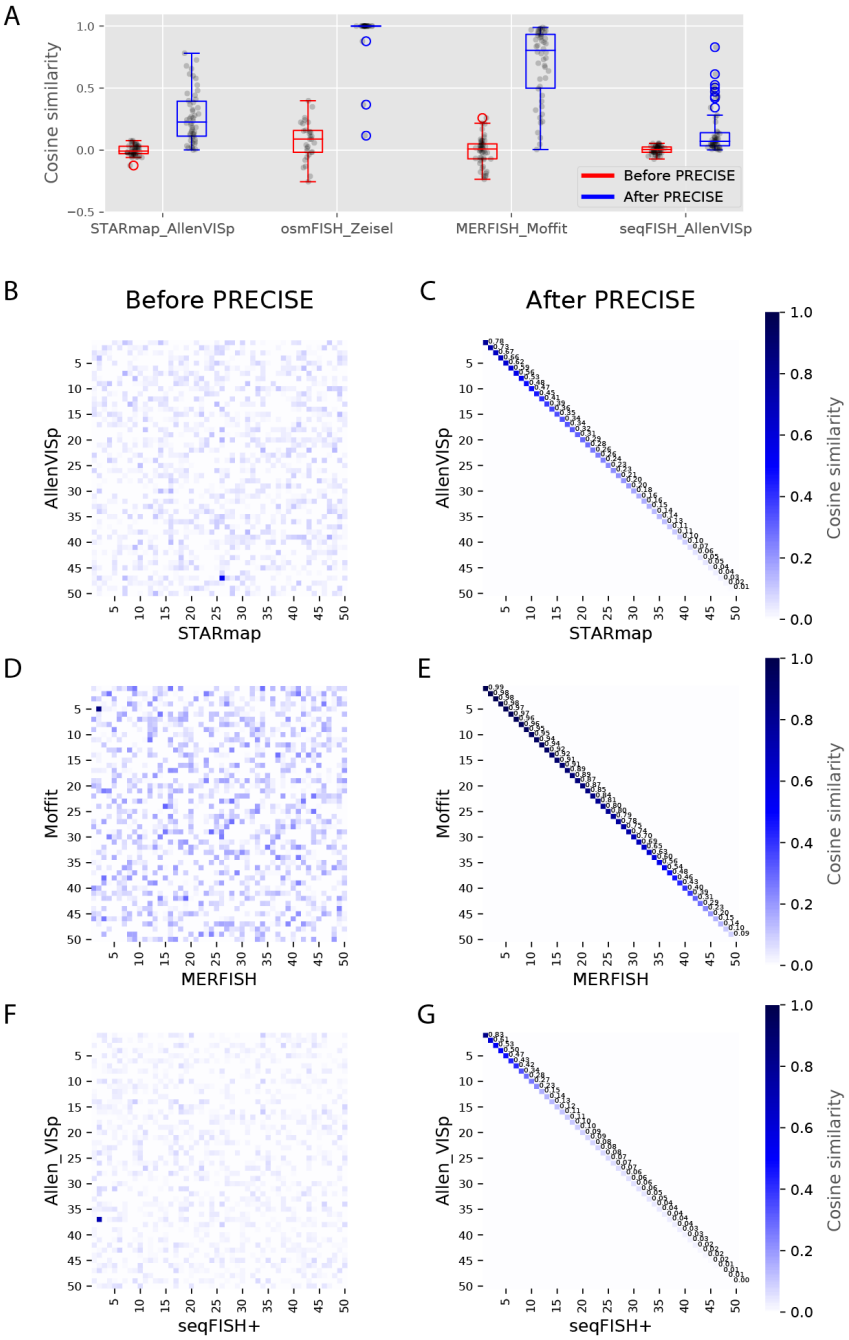
The implementation code of SpaGE, as well as the benchmarking code, is available in the GitHub repository, at https://github.com/tabdelaal/SpaGE. The code is released under MIT license. All datasets used are publicly available data, for convenience datasets can be downloaded from Zenodo (https://doi.org/10.5281/zenodo.3967291).

BIBLIOGRAPHY

1. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172,** 1091–1107.e17 (2018).
2. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566,** 496–502 (2019).
3. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174,** 999–1014.e22 (2018).
4. Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science (80-.).* **362,** (2018).
5. Codeluppi, S. *et al.* Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods* **15,** 932–935 (2018).
6. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (80-.).* **348,** (2015).
7. Eng, C. H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568,** 235–239 (2019).
8. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science (80-.).* **361,** eaat5691 (2018).
9. Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science (80-.).* **363,** 1463–1467 (2019).
10. Xia, C., Fan, J., Emanuel, G., Hao, J. & Zhuang, X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 19490–19499 (2019).
11. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nature Reviews Genetics* **20,** 257–272 (2019).
12. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33,** 495–502 (2015).
13. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33,** 503–509 (2015).
14. Nitzan, M., Karaiskos, N., Friedman, N. & Rajewsky, N. Gene expression cartography. *Nature* **576,** 132–137 (2019).
15. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888–1902.e21 (2019).
16. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177,** 1873–1887.e17 (2019).
17. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16,** 1289–1296 (2019).
18. Lopez, R. *et al.* A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. *arXiv* (2019).
19. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21,** 12 (2020).
20. Mourragui, S., Loog, M., Van De Wiel, M. A., Reinders, M. J. T. & Wessels, L. F. A. PRECISE: A domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. in *Bioinformatics* **35,** i510–i519 (2019).
21. Gene H. Golub & Van Loan, C. F. *Matrix Computations, 4th Edition*. *The Johns Hopkins*

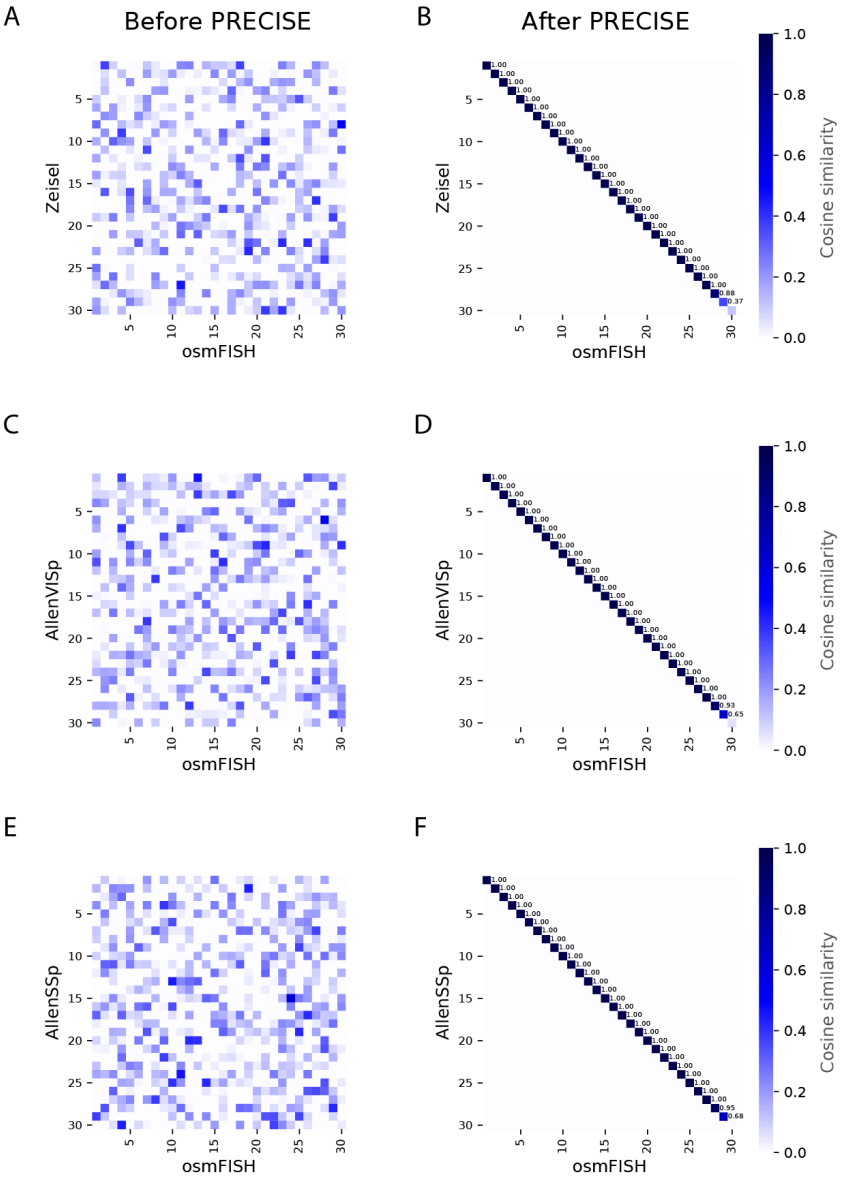
University Press (2013).
22. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563,** 72–78 (2018).
23. Chatterjee, S. *et al.* Nontoxic, double-deletion-mutant rabies viral vectors for retrograde targeting of projection neurons. *Nat. Neurosci.* **21,** 638–646 (2018).
24. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science (80-.).* **347,** 1138–1142 (2015).
25. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17,** 261–272 (2020).
26. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15,** 1053–1058 (2018).
27. Li, H., Calder, C. A. & Cressie, N. Beyond Moran's I: Testing for spatial dependence based on the spatial autoregressive model. *Geogr. Anal.* **39,** 357–375 (2007).

SUPPLEMENTARY MATERIALS

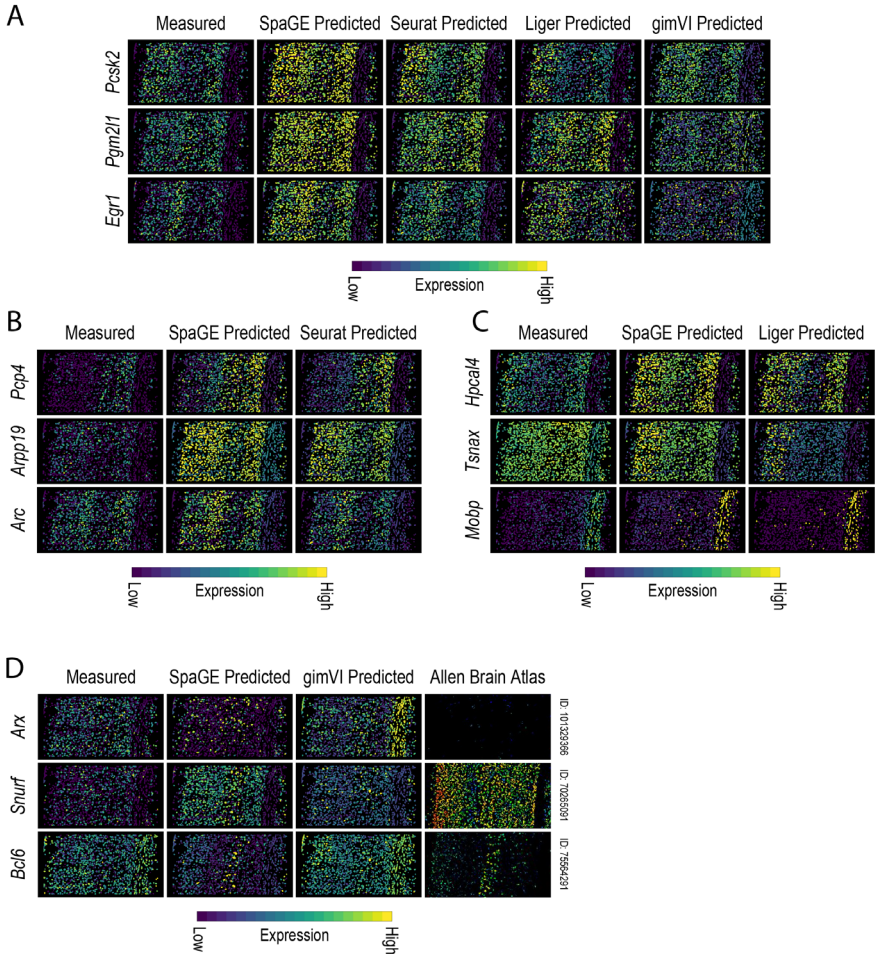


Supplementary Figure 8.1 Pairwise cosine similarity matrices before and after PRECISE. (A) Boxplots showing the diagonal (one-to-one) Cosine similarity between the independent *PCs* (before PRECISE) of both datasets in each dataset pair, and between the *PVs* (after PRECISE). **(B,D,F)** Pairwise

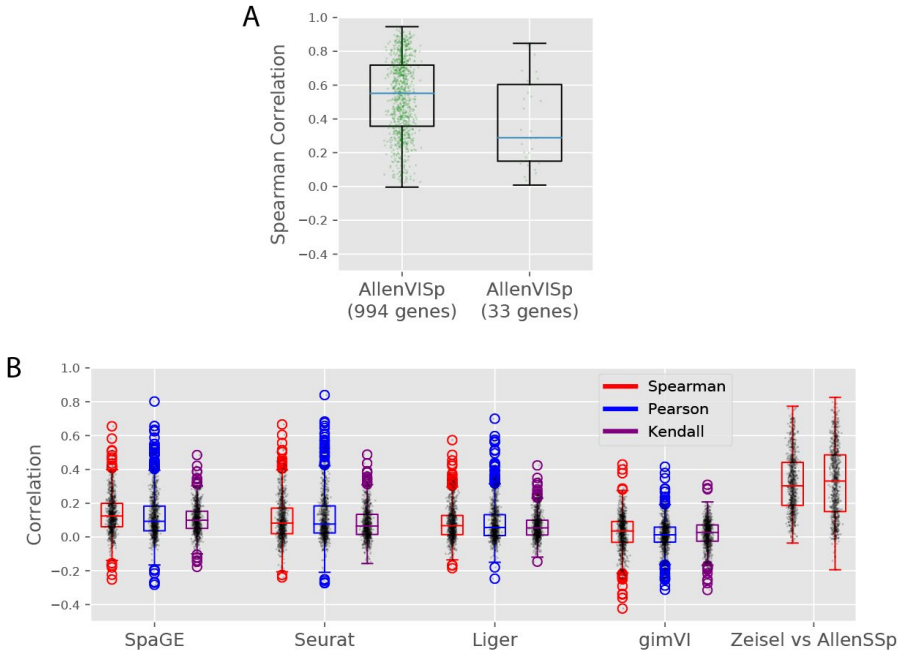
cosine similarity matrices between the *PCs* (before PRECISE) of the **(B) STARmap_AllenVISp**, **(D) MERFISH_Moffit**, and **(F) seqFISH_AllenVISp** dataset pairs, showing no one-to-one correspondence. **(C,E,G)** Pairwise cosine similarity matrices between the *PVs* (after PRECISE) of the **(C) STARmap_AllenVISp**, **(E) MERFISH_Moffit**, and **(G) seqFISH_AllenVISp** dataset pairs, showing a clear one-to-one diagonal similarity.



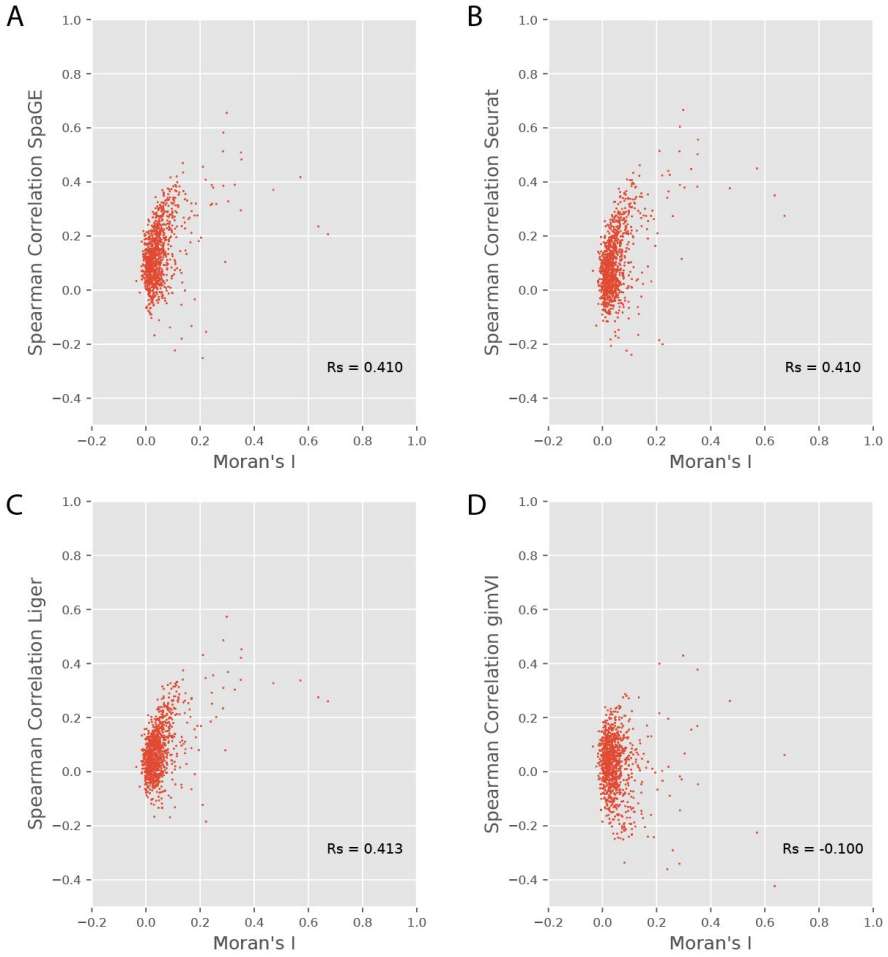
Supplementary Figure 8.2 Pairwise cosine similarity matrices before and after PRECISE. (A,C,E) Pairwise cosine similarity matrices between the *PCs* (before PRECISE) of the **(A) osmFISH_Zeisel**, the **(C) osmFISH_AllenVISp** and the **(E) osmFISH_AllenSSp** dataset pairs, showing no one-to-one correspondence. **(B,D,F)** Pairwise cosine similarity matrices between the *PVs* (after PRECISE) of the **(B) osmFISH_Zeisel**, the **(D) osmFISH_AllenVISp** and the **(F) osmFISH_AllenSSp** dataset pairs, showing a clear one-to-one diagonal similarity.



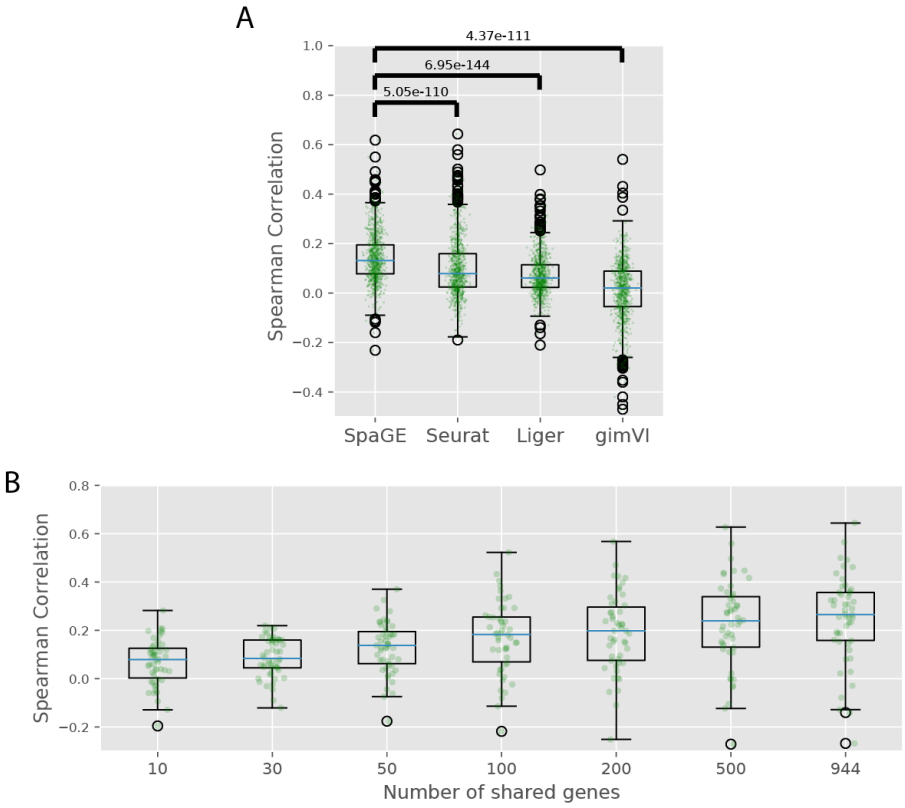
Supplementary Figure 8.3 Top predicted genes of each method using the STARmap_AllenVISp dataset pair. (A) Comparison of the top 3 genes predicted by SpaGE. Each row corresponds to a single gene, first column from the left shows the measured spatial gene expression in the **STARmap** dataset, while other columns show the corresponding predicted expression pattern by SpaGE, Seurat, Liger and gimVI. **(B-D)** Comparison of the top 3 genes predicted by **(B)** Seurat, **(C)** Liger and **(D)** gimVI, excluding the top 10 predicted genes by SpaGE. Each row corresponds to a single gene, first column from the left shows the measured spatial gene expression in the **STARmap** dataset, the second column shows the corresponding predicted expression pattern by SpaGE, while the third column shows predicted expression pattern by **(B)** Seurat, **(C)** Liger and **(D)** gimVI. All predictions were obtained using the leave-one-gene-out cross validation experiment. In **(D)**, the fourth (additional) column shows the Allen Brain Atlas spatial expression, showing more visual agreement with the predicted spatial patterns using SpaGE.



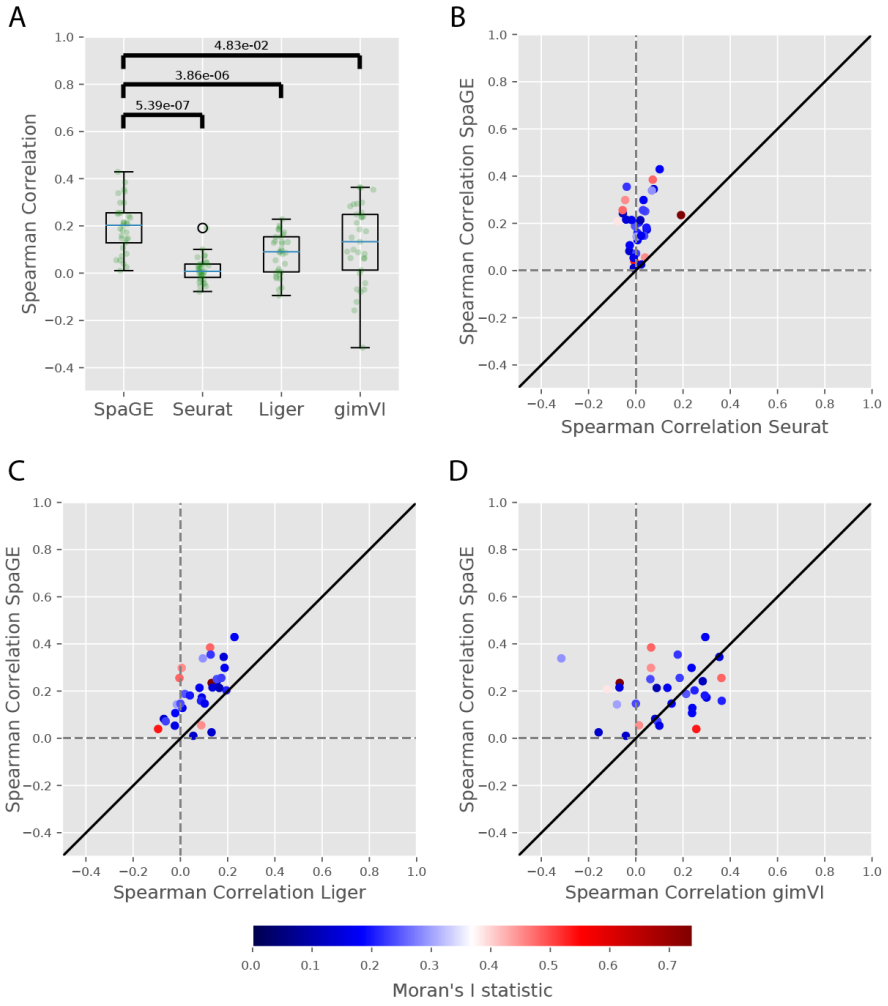
Supplementary Figure 8.4 (A) Boxplots showing the kNN regression prediction performance when applied on the AllenVISp, using 994 and 33 genes of the STARmap and osmFISH datasets, respectively. The green dots show the correlation values for individual genes. **(B)** Boxplots showing different correlation measures for all methods using the **STARmap_AllenVISp** dataset pair. The right most boxplot pair shows the performance of SpaGE when applied on two scRNA-seq datasets, **Zeisel** and **AllenSSp**. The black dots show the correlation values for individual genes.



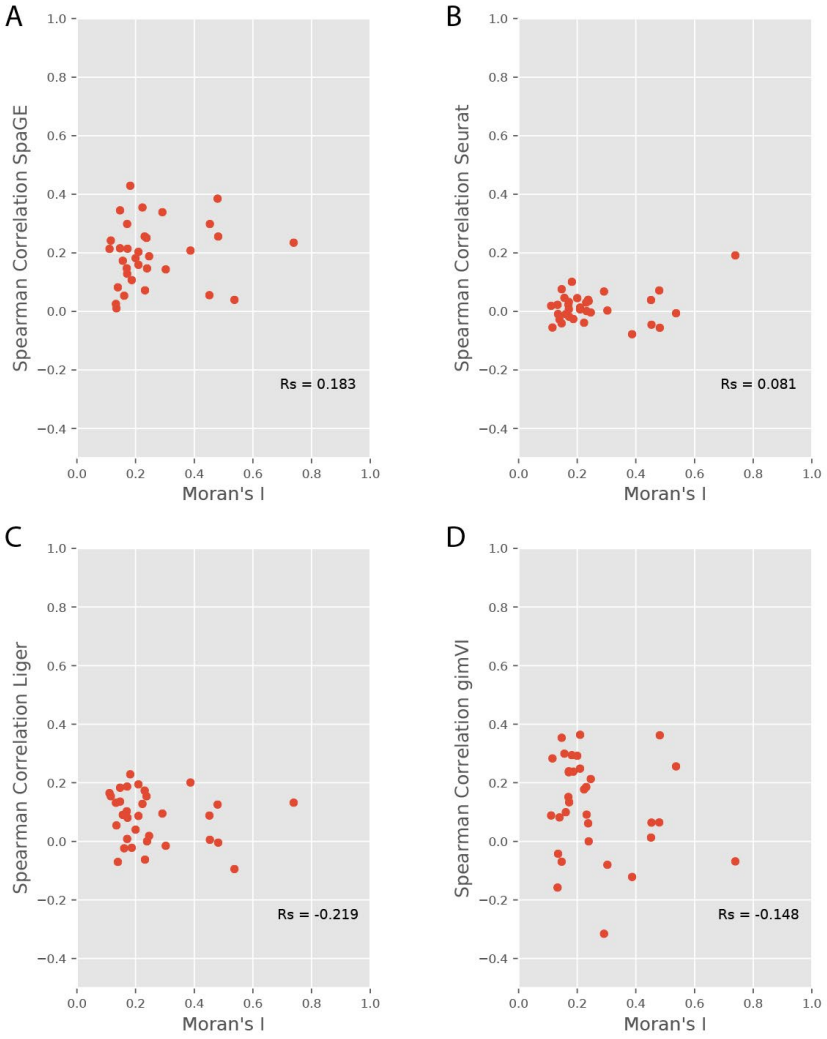
Supplementary Figure 8.5 (A-D) Scatter plots showing the relation between the Moran's I statistic and the prediction correlation of each gene, using the **STARmap_AllenVISp** dataset pair. Moran's I (x-axis) are calculated using the STARmap dataset and prediction correlation values (y-axis) were obtained by **(A)** SpaGE, **(B)** Seurat, **(C)** Liger and **(D)** gimVI. The Rs values correspond to the Spearman Rank correlation between the Moran's I statistic and the prediction performance of each method.



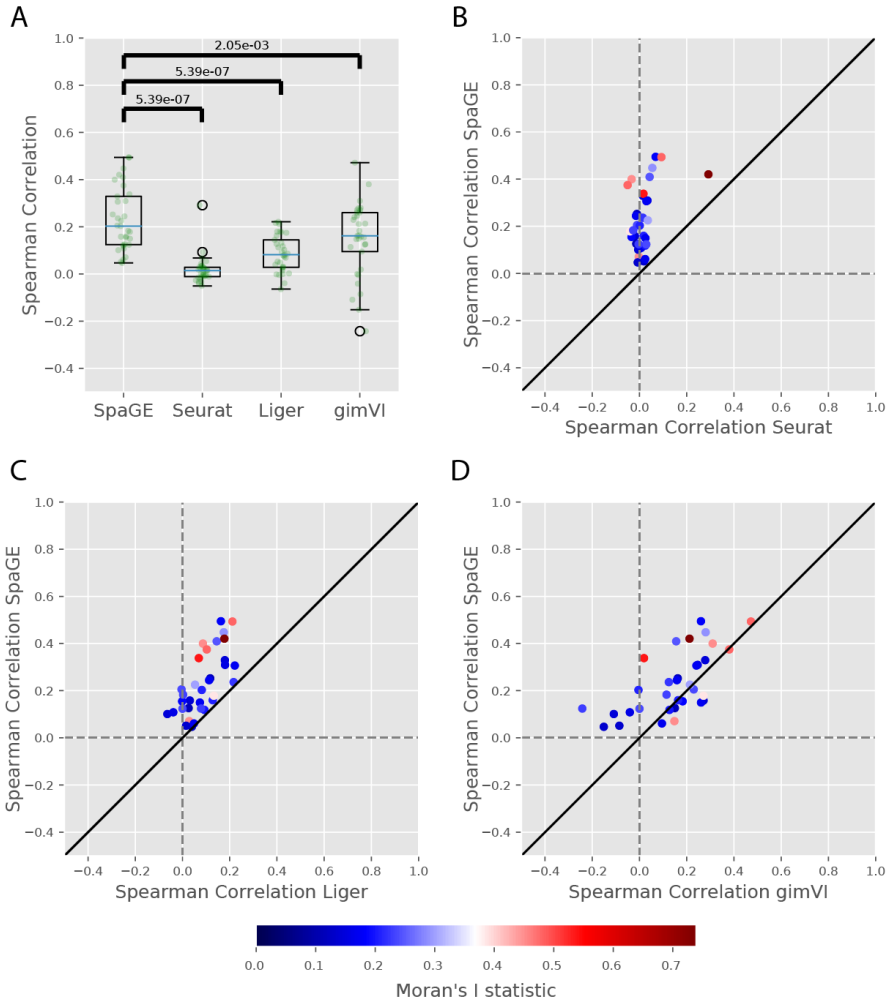
Supplementary Figure 8.6 (A) Boxplots showing the Spearman correlations of each method when excluding the 100 most correlated genes with the left-out gene from the shared gene set. The blue lines show the median correlation across all genes with a better performance for SpaGE. The green dots show the correlation values for individual genes. The p-values show the significant difference between all correlation values of SpaGE and each other method, using a paired Wilcoxon rank-sum test. **(B)** Boxplots showing the prediction performance of a test set of 50 genes, in terms of Spearman Rank correlations, using downsampled sets of 10, 30, 50, 100, 200 and 500 shared genes compared to using all 944 genes in the **STARmap_AllenVISp** dataset pair.



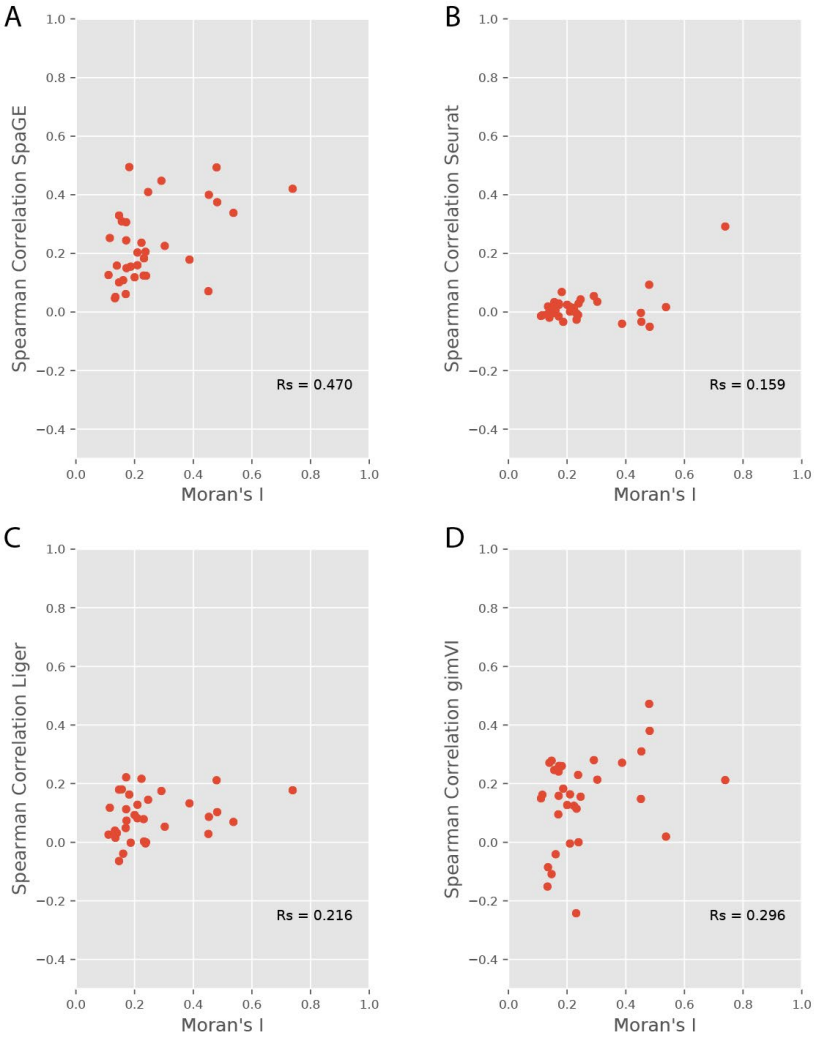
Supplementary Figure 8.7 Prediction performance comparison for the osmFISH_Zeisel dataset pair. (A) Boxplots showing the Spearman correlations for the leave-one-gene-out cross validation experiment for each method. The blue lines show the median correlation across all genes with a better performance for SpaGE. The green dots show the correlation values for individual genes. The p-values show the significant difference between all correlation values of SpaGE and each other method, using a paired Wilcoxon rank-sum test. **(B-D)** Detailed performance comparison between SpaGE and **(B)** Seurat, **(C)** Liger, **(D)** gimVI. These scatter plots show the correlation value of each gene across two methods. The solid black line is the y=x line, the dashed lines show the zero correlation. Points are colored according to the Moran's I statistic of each gene. All scatter plots show that the majority of the genes are skewed above the y=x line, showing an overall better performance of SpaGE over other methods.



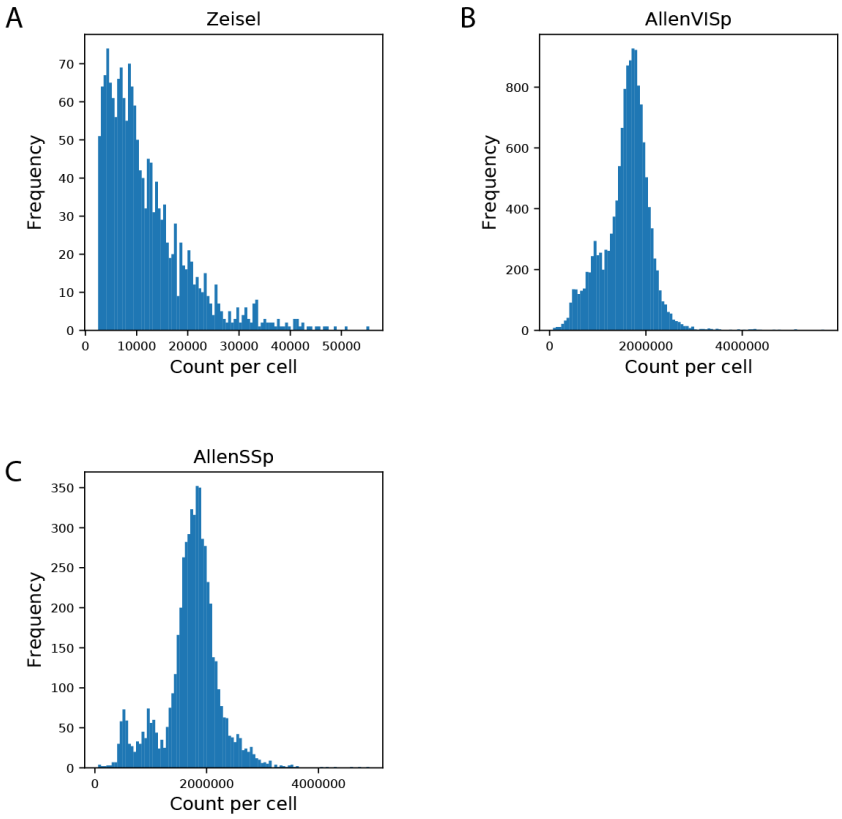
Supplementary Figure 8.8 (A-D) Scatter plots showing the relation between the Moran's I statistic and the prediction correlation of each gene, using the **osmFISH_Zeisel** dataset pair. Moran's I (x-axis) are calculated using the **osmFISH** dataset and prediction correlation values (y-axis) were obtained by **(A)** SpaGE, **(B)** Seurat, **(C)** Liger and **(D)** gimVI. The Rs values correspond to the Spearman Rank correlation between the Moran's I statistic and the prediction performance of each method.



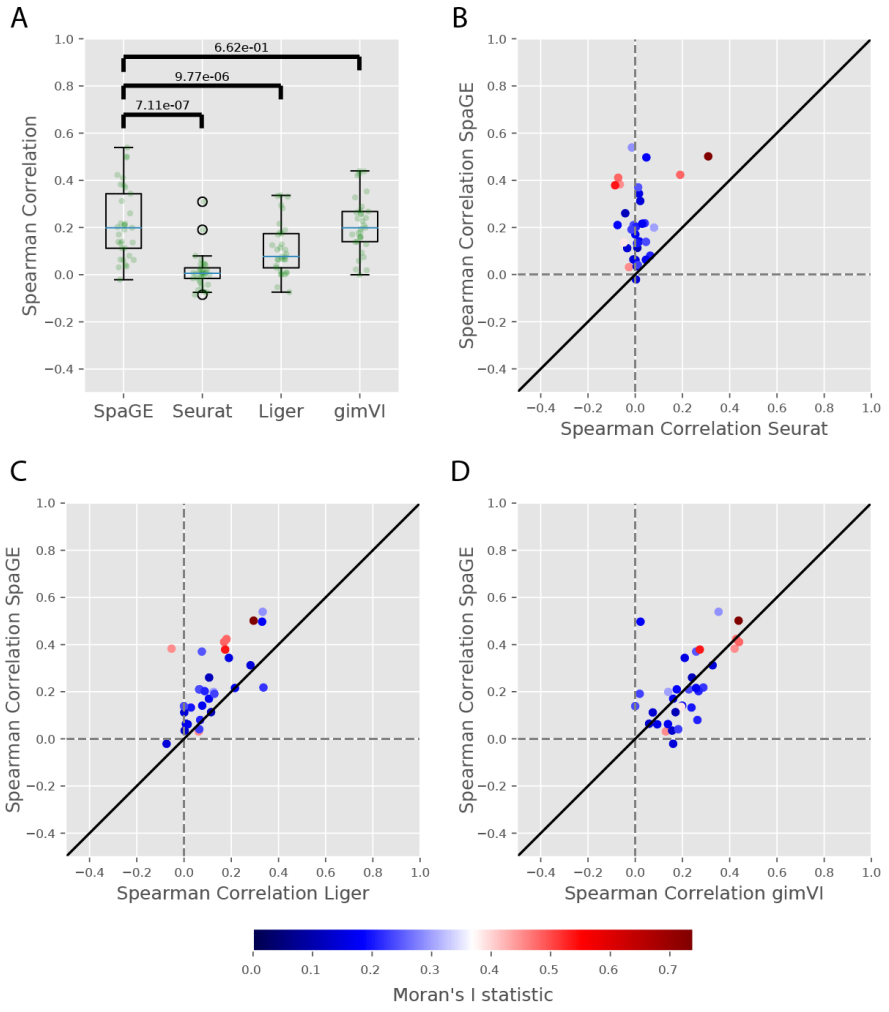
Supplementary Figure 8.9 Prediction performance comparison for the osmFISH_AllenVISp dataset pair. (A) Boxplots showing the Spearman correlations for the leave-one-gene-out cross validation experiment for each method. The blue lines show the median correlation across all genes with a better performance for SpaGE. The green dots show the correlation values for individual genes. The p-values show the significant difference between all correlation values of SpaGE and each other method, using a paired Wilcoxon rank-sum test. **(B-D)** Detailed performance comparison between SpaGE and **(B)** Seurat, **(C)** Liger, **(D)** gimVI. These scatter plots show the correlation value of each gene across two methods. The solid black line is the y=x line, the dashed lines show the zero correlation. Points are colored according to the Moran's I statistic of each gene. All scatter plots show that the majority of the genes are skewed above the y=x line, showing an overall better performance of SpaGE over other methods.



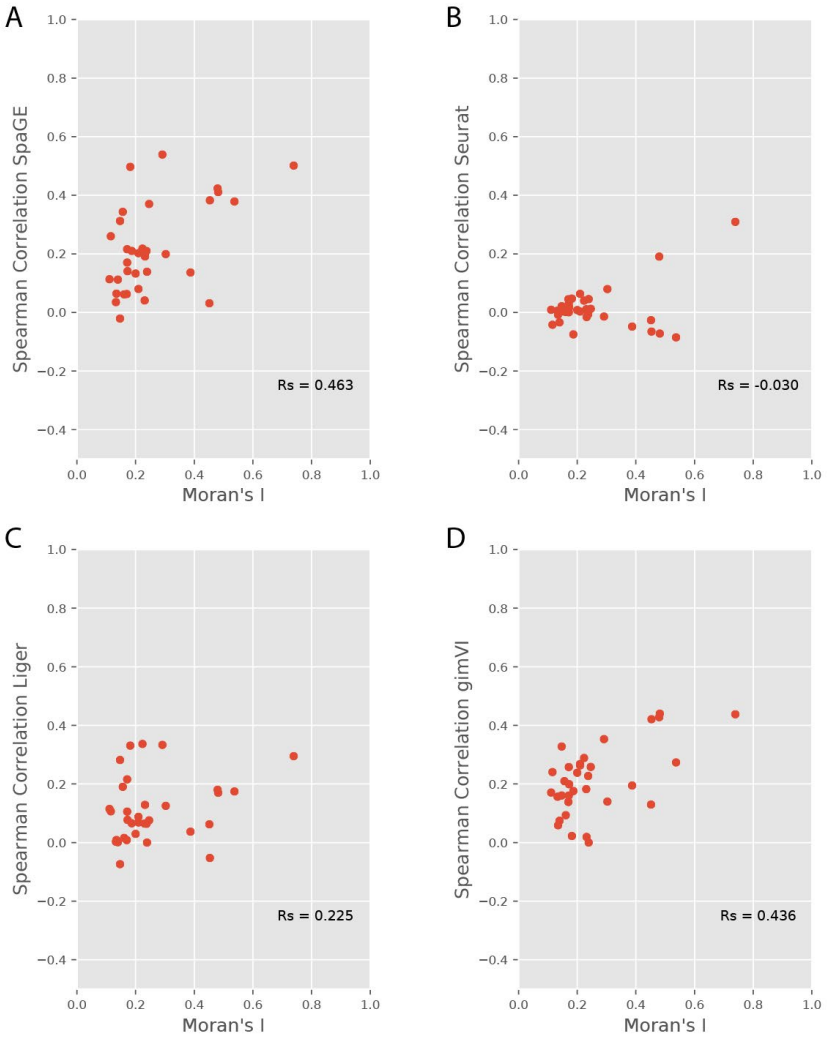
Supplementary Figure 8.10 (A-D) Scatter plots showing the relation between the Moran's I statistic and the prediction correlation of each gene, using the **osmFISH_AllenVISp** dataset pair. Moran's I (x-axis) are calculated using the **osmFISH** dataset and prediction correlation values (y-axis) were obtained by **(A)** SpaGE, **(B)** Seurat, **(C)** Liger and **(D)** gimVI. The Rs values correspond to the Spearman Rank correlation between the Moran's I statistic and the prediction performance of each method.



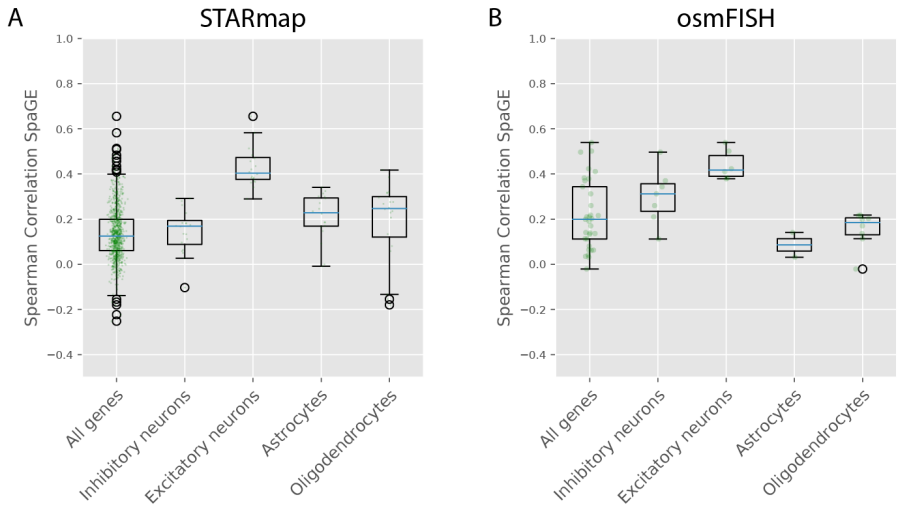
Supplementary Figure 8.11 Sequencing depth across difference scRNA-seq reference datasets. Histograms showing the distribution of the RNA count per cell of the **(A) Zeisel**, **(B) AllenVISp**, and **(C) AllenSSp** datasets, respectively. The **AllenVISp** and **AllenSSp** datasets have comparable sequencing depths, while the sequencing depth of the **Zeisel** dataset is ~140x lower.



Supplementary Figure 8.12 Prediction performance comparison for the osmFISH_AllenSSp dataset pair. (A) Boxplots showing the Spearman correlations for the leave-one-gene-out cross validation experiment for each method. The blue lines show the median correlation across all genes with a better performance for SpaGE. The green dots show the correlation values for individual genes. The p-values show the significant difference between all correlation values of SpaGE and each other method, using a paired Wilcoxon rank-sum test. **(B-D)** Detailed performance comparison between SpaGE and **(B)** Seurat, **(C)** Liger, **(D)** gimVI. These scatter plots show the correlation value of each gene across two methods. The solid black line is the y=x line, the dashed lines show the zero correlation. Points are colored according to the Moran's I statistic of each gene. All scatter plots show that the majority of the genes are skewed above the y=x line, showing an overall better performance of SpaGE over other methods.



Supplementary Figure 8.13 (A-D) Scatter plots showing the relation between the Moran's I statistic and the prediction correlation of each gene, using the **osmFISH_AllenSSp** dataset pair. Moran's I (x-axis) are calculated using the **osmFISH** dataset and prediction correlation values (y-axis) were obtained by **(A)** SpaGE, **(B)** Seurat, **(C)** Liger and **(D)** gimVI. The Rs values correspond to the Spearman Rank correlation between the Moran's I statistic and the prediction performance of each method.



Supplementary Figure 8.14 (A,B) Boxplots showing the prediction performance of SpaGE for cell type marker genes compared to the overall performance across all genes, using **(A)** the **STARmap** and **(B)** the **osmFISH** datasets. The green dots show the correlation values for individual genes.

Supplementary Table 8.1 PRECISE results summary. For each dataset pair, we summarize the initial and final number of principal vectors (PVs), and the percentage of explained variance by the final PVs for each dataset.

Spatial_scRNA-seq dataset pair	**Initial # of PVs (d)**	**Final # of PVs (d')**	**Explained variance spatial**	**Explained variance scRNA-seq**
STARmap_AllenVISp	50	19	9.97%	34.83%
osmFISH_Zeisel	30	29	94.25%	95.26%
osmFISH_AllenSSp	30	29	94.23%	97.49%
osmFISH_AllenVISp	30	29	94.39%	96.90%
MERFISH_Moffit	50	42	52.34%	45.36%
seqFISH_AllenVISp	50	8	5.96%	11.77%

CHAPTER 9

DISCUSSION

Single-cell technologies became essential to understand the cellular composition and organization within a specific tissue. The rapid advances of single-cell technologies generating high-dimensional large scale datasets poses several challenges in the data analysis. Over the past decade, numerous computational tools have been developed for single-cell data analysis, and have been successfully applied to answer challenging biological questions, such as the generation of a cellular atlas of cancerous tissue[1], or the identification of biomarkers for cancer progression[2]. In this thesis, we introduced a set of computational tools developed to address several challenges in the data analysis. In the following sections, we discuss our main conclusions and propose future extensions to our work regarding the three main challenges that we identified in the introduction: interaction, identification and integration of single-cell data. Finally, we briefly discuss our view on the future of single-cell analysis in terms of the technological advances, new challenges and opportunities arising, and the need for data analysis standardization.

9.1 INTERACTION

In **Chapter 2**, we introduced SCHNEL, an upgraded version of the graph-based clustering method scalable to datasets having millions of cells. To make the clustering computationally feasible, SCHNEL applies the Louvain community detection algorithm to a downsampled sets of cells. These sets are selected to be representative of the full data structure using the HSNE paradigm. SCHNEL showed robust clustering across different single-cell datasets including protein (cytometry) and gene (scRNA-seq) expression measurements, suggesting SCHNEL to be a general clustering tool for single-cell data.

To test this generalization even further, we can test the applicability of SCHNEL to other single-cell modalities such as scATAC-seq, measuring the chromatin accessibility at the single-cell resolution. In general, directly applying analysis methods designed for cytometry and scRNA-seq data on scATAC-seq data may not produce proper results, due to the increased sparsity of the data and the near-binary nature. Imputation methods are often involved in the analysis of scATAC-seq data, prior to the downstream analysis, to enrich the data and reduce the sparsity[3,4]. SCHNEL is designed and implemented in a modular way, sequentially connecting two powerful analysis modules (HSNE and Louvain). Consequently, adding an imputation module prior to the HSNE can generalize SCHNEL to be also applicable scATAC-seq data.

In addition, we showed that SCHNEL can provide clusterings for all scales of the hierarchy generated by HSNE. This provides different levels of details in the data and permits the user to interactively pick the most suitable clustering detecting the cell types or states of interest. However, in some cases, a complete automated analysis is required without any user input. Thus, further improvement can be carried out to define the optimal clustering and provide a single clustering result. A quantitative unsupervised score (e.g. silhouette score) can be used to evaluate the clustering at each scale, and guide the selection of the best clustering achieving the highest reparability between the different cell clusters.

In **Chapter 3**, we presented Cytosplore-transcriptomics, a complete platform to analyze scRNA-seq data, including data preprocessing, data visualization and downstream analysis. The main theme of Cytosplore-transcriptomics is to provide interactive analysis based on visual exploration of the data using low-dimensional visualizations. Using HSNE, Cytosplore-transcriptomics is scalable to large datasets having millions of cells, and can interactively produce low-dimensional visualizations of the data hierarchy providing different levels of details.

Scalability is a crucial feature, as the amount of single-cell data has been exponentially growing through the past decade[5], and it will continue growing with more technological advances. Not only the amount of cells is increasing, but the number of defined cell populations as well. At some point, building hierarchies and finding representative cells cannot be properly extended, as some populations will not have enough representative cells at the scale which is computationally feasible to analysis. Therefore, continuous efforts are required in order to improve the scalability of the data analysis methods even further, eventually reaching a linear computational complexity[6,7].

Visualization of high-dimensional single-cell data into a two-dimensional embedding has being widely used to assess the cellular composition (cell types), overlaying metadata or the expression of specific genes/proteins[8]. This assessment can help studying many cellular aspects including differentiation trajectories[9,10], and compositional analysis across different conditions[11]. These two-dimensional maps, such as tSNE[12] and UMAP[13], are usually constructed in a complete unsupervised manner, such that the distances in the XY coordinates are comparable to the distances in the high-dimensional space. It would be interesting to explore possibilities to embed the cells using a supervised or semi-supervised approach, such that cells with similar identities (labels) can be enforced to be closer in the resulting map. This supervised guidance is of great use when visualizing data across different batches[14], creating a batch corrected low-dimensional map which is important for further downstream analysis.

9.2 IDENTIFICATION

In **Chapter 4** and **Chapter 5**, we evaluated replacing the clustering methods with classification methods for cell type identification in scRNA-seq and cytometry data. We showed that linear models such as Linear SVM and LDA performed well for scRNA-seq and mass cytometry, respectively. Further, these linear classifiers outperformed complex non-linear machine learning and deep learning methods. These classification methods make use of the large amount of labeled data available nowadays, and offer the opportunity for automated and reproducible cell identification.

Generally, a good classification model mostly depends on a good training data. Ideally, an atlas containing all possible cell populations of a certain tissue would represent an optimal training data. Recent efforts have been made to generate such atlases for certain species. The Human Cell Atlas (HCA) consortium, started in 2017, aims to generate a comprehensive map of all human cells across different tissues using a variety of single-cell technologies[15]. Meanwhile, others studies focused on generating a Mouse Cell Atlas (MCA) using scRNA-seq[16,17].

Despite these efforts, a complete atlas is currently still lacking and requires years of data collection and analysis. Further, such atlas, containing all possible cellular subpopulations covering a huge number of diseases, may never exists. One alternative is to map the annotated single-cell populations to cell ontology terms, and use this mapping to train a classification model. This opens the possibility to label cells with new annotation not present in the single-cell dataset at hand[18]. However, these cell ontologies are not developed for single-cell data, and newly discovered subpopulations might not map correctly to the cell ontology terms.

Another alternative is to build a classifier from multiple annotated datasets covering specific tissue. Such classifier should match different cell populations across studies, and combine

this information in some form (e.g. using a hierarchy) that can be continuously extended whenever a new cell population is defined[19]. However, this process in not straightforward due to the inconsistent terminology used to name different cell populations. Furthermore, this can be also applied on a smaller scale within one study. For example, considering one large cohort study with hundreds of individuals, the first replicates might be considered as training data. These replicates should be chosen to represent the full dataset. In a study containing samples with different biological conditions, for example a case-control study, the training data must contain enough and equivalent samples from all conditions.

An important aspect of using classification for cell type identification is incorporating a rejection option. The classifier should be able to flag new cells, not present in the training data, as "unknown" to avoid forcing a misclassification to a wrong cell class. One approach is to reject cells having a prediction probability below a certain threshold, which may be considered as confidence threshold on the predicted cell identity. We showed that such approach did not properly work in all cases, especially when there is a single population in the training data which is similar to the new cell to be predicted. This can be improved by adding another distance based threshold, where a cell should be also rejected if the distance between that cell and the predicted class is above a certain threshold[20]. Another improvement is to apply a classifier with a tight decision boundary, like the one-class SVM[19]. This can indeed reject dissimilar cells and decrease the false positives, but on the other hand, it might also increase the false negatives producing an overall lower prediction performance.

When combining multiple datasets in order to create a cell atlas, or when using such atlas to classify new cells, technical differences between datasets represent a major problem that might completely skew the cell annotations. These technical differences, often called batch effects, can result from different experimental protocols and different machines used to generate the data within the same lab or across different labs. Batch correction methods must be applied to remove these technical differences between multiple datasets used to train the classifier, and between the training (reference) and testing (query) datasets before producing predictions for the query data. The latter can be applied in a domain adaptation manner, where the query data is mapped to the reference data in order to apply the pre-trained classifier. A novel deep learning data integration method, scArches[21], proposed using transfer learning and parameter optimization to build efficient reference models. Pre-trained reference models can be shared without the need of raw data. Additional datasets can be iteratively integrated to update the reference model, while query data can be integrated and annotated using such model.

9.3 INTEGRATION

In this thesis, we introduced two data integration methods to enhance and extend the number of measured features beyond the current technology limitations. In **Chapter 7**, we introduced CyTOFmerge, integrating different mass cytometry datasets measured from the same biological sample, resulting in an extended number of proteins markers per cell. Downstream analysis of the integrated data further reveals the cellular heterogeneity by defining new cell subpopulations. While in **Chapter 8**, we introduced SpaGE, integrating two single-cell modalities, scRNA-seq and spatial transcriptomics. SpaGE produced whole transcriptome spatial data showing correct in-silico spatial expression patterns of genes not originally measured in the spatial data.

To perform accurate integration, SpaGE relied in its core on the domain adaption method PRECISE[22]. PRECISE was used to eliminate the technical differences between datasets and

produce aligned datasets, where a simple kNN regression was enough to correctly estimate the spatial gene expressions. Building on that, PRECISE can serve as a general data integration method, which can be used to correct for batch effects within one datasets, or to perform data integration between different single-cell modalities. This has a large potential, as PRECISE is a linear model which is easily scalable to large datasets having millions of cells. As shown within SpaGE, time and memory requirements are much lower than the current state-of-the-art data integration methods. However, one limitation is that currently PRECISE only works for a pair of datasets (reference and query). While in most cases, multiple batches or datasets are integrated to perform one analysis. A solution could be to apply PRECISE iteratively, integrating two batches/datasets first and use the integrated version as new reference to further add more batches/datasets.

Currently, most data integration methods rely on a common set of features to perform the integration. However, in some cases, single-cell multi-omics datasets are measured from the same tissue, with unpaired cells and unmatched features as well, which makes the data integration task more challenging. Even with matching features, it's not always straightforward to have a clear one-to-one matching between features. For example, when integrating two scRNA-seq datasets from human and mouse, not all human and mouse genes are one-to-one matched. Recently, data integration methods relied on unsupervised manifold alignment to integrated datasets with no correspondence between cells or features[23,24]. First, a low-dimensional embedding is obtained for each single-cell dataset separately. Next, a common low-dimensional embedding, having the aligned cells, is produced by aligning the distributions of these separate low-dimensional embeddings across datasets.

Applying this manifold alignment in a complete unsupervised manner might produce wrong matching between cells. Alternatively, cells in each dataset can be analyzed and annotated separately, cell identity labels are then matched across datasets. Next, the manifold alignment of the low-dimensional embeddings is applied in a semi-supervised manner using a small proportion of the labeled cells across datasets[25]. Although this procedure requires additional information (cell identity labels), this procedure prevents wrong matching of similar cells. Additionally, cell populations present in only one dataset with no matching cells in the other dataset, are kept separate in the final aligned embedding avoiding forced mismatching.

Further, data integration might increase the ability to study the dynamics of cellular differentiation. Several computational methods aim to infer cellular dynamics information from the static snapshot data captured using scRNA-seq[9,10]. These methods define a differentiation trajectory between different cell populations obtained the pool of cells. Fluorescence-based live cell imaging provides a better view of the cellular dynamic transition between different states, such information can be missed in snapshot data[26]. However, these live cell imaging techniques are limited in the number of dynamic features (genes) that can be measured. Integrating snapshot (scRNA-seq or spatial transcriptomics) data with fluorescence-based live imaging data will provide a continuous dynamic measurement of a large number of features, thus improving the study of cellular dynamics and lineage differentiation.

9.4 PERSPECTIVE ON THE FUTURE OF SINGLE-CELL ANALYSIS

Current single-cell technologies can measure several molecular features including genomes, epigenomes, transcriptomes, proteomes and spatial localization. Separately, each data type provides a different view of the cellular state, and, when combined, can resolve complex biological processes (e.g. gene-regulatory networks)[27]. Single-cell multi-omics (2019 method

of the year[28]) represents the latest advance in the single-cell field, where multiple molecular features are measured simultaneously from the same cell[29]. Some methods can measured two data types, including CITE-seq[30] (transcriptome and proteome), sci-CAR[31] and SNARE-seq[32] (transcriptome and chromatin accessibility), while scNMT[33] can even measure three data types simultaneously (transcriptome, methylome and chromatin accessibility).

These single-cell multi-omics technologies provide new opportunities and challenges for the data analysis. As simultaneous features are measured from the same cells, we can study and model specific mechanisms connecting these different features. For example, CITE-seq data can be used to model the dynamic translation process of mRNA to proteins. These models can be learned separately across different cell types, and might reveal cell-type specific regulation.

Although measuring multiple omics from the same cell provides a better view of the cellular identity, grouping the cells into different populations based on these multi-omics is more complicated. Combined clustering analysis should be performed to account for several data types together. One way is to cluster every data type separately, and continue to find one overall fine-grained clustering by combining different clustering results from each data type (late integration). Alternatively, features from all data types can be used to perform a single clustering analysis (early integration). Although this will directly produce one clustering, variation sources across data types should be equally weighted, otherwise the data type with large variation will overcome the clustering result.

Supervised learning approaches can benefit from multi-omics datasets as well. The classification model can then be trained with multiple data types, allowing to automatically annotate cells having one or more measured data types. The distance metrics should however fit each data type separately, which could be learned (metric learning), with multiple kernel-based similarity learning[34] being just one example.

Further, generating low-dimensional visualization maps of multiple data types will enrich the interactive analysis. Separate maps can be obtained for each data type showing multiple views of the same cell. As these are paired cells, it is interesting to compare the local structure around each cell across different data types, showing how different the cellular composition and interaction across different single-cell modalities. Additionally, an important area of research is to combine single-cell multi-omics data with spatial information of the cells. In various tissues, spatial distribution is a key determinant of the cellular identity. Similar to SpaGE, various types of molecular features can be overlaid on the spatial localization of the cells. This will help to further study the cellular structure and cell-cell communication.

Single-cell technology development had been rapidly advancing over the past decade. Several techniques are available nowadays to measure various molecular features. This resulted in a huge number of studies applying single-cell techniques with a large number of biological findings (e.g. discovery of a new cell subpopulation associated with specific disorder). Although single-cell analysis has been effectively applied to various biological fields, like brain and cancer research, most of these new discoveries are not well enough reproduced across different studies from different labs, affecting the overall validation. To further transfer these new discoveries into clinical applications, more cohort single-cell studies[35] are needed in the future to reproduce and validate these new discoveries, and strengthen their biological impact.

Parallel to the fast growing of single-cell technology, a large number of data analysis methods have been specifically developed for single-cell data. On the one hand, this offers the users community alternative solutions for similar technical problems arising during data analysis. But on the other hand, this leaves users unsure which method fits best their data. More importantly, this affects the reproducibility issue as the analysis of the same data may have different outcomes when applying different methods, leading to alternative interpretations of the same data. Recently, large efforts have been made in benchmarking studies comparing the outcomes of different analysis methods for specific technical problems[36–39]. These benchmarking studies are essential to be performed on a continuous time scale, to eventually converge towards a standardized analysis of single-cell data.

9.5 CONCLUDING REMARKS

In this thesis, we have introduced several computational methods to aid and improve the analysis of single-cell data. Several methods have been used, including hierarchical representation of the data to improve the scalability towards large datasets. Supervised learning has been used to substitute unsupervised learning for automatic cell identification. Data integration, combined with domain adaptation, has been used to enrich the data beyond the current technical limitations by extending the number of molecular features per single-cell. Together, these methods improve the interpretation of the data, and guide the future computational development for single-cell analysis.

BIBLIOGRAPHY

1. Chevrier, S. *et al.* An Immune Atlas of Clear Cell Renal Cell Carcinoma. *Cell* **169,** 736–749 (2017).
2. Cortese, N., Carriero, R., Laghi, L., Mantovani, A. & Marchesi, F. Prognostic significance of tumor-associated macrophages: past, present and future. *Seminars in Immunology* (2020). doi:10.1016/j.smim.2020.101408
3. Xiong, L. *et al.* SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat. Commun.* **10,** (2019).
4. Albrecht, S., Andreani, T., Andrade-Navarro, M. & Fontaine, J.-F. Single-cell ChIP-seq imputation with SIMPA by leveraging bulk ENCODE data. *bioRxiv* 1–15 (2019). doi:10.1101/2019.12.20.883983
5. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols* **13,** 599–604 (2018).
6. Pezzotti, N. *et al.* GPGPU Linear Complexity t-SNE Optimization. *IEEE Trans. Vis. Comput. Graph.* **26,** 1172–1181 (2020).
7. Bodenheimer, T. *et al.* FastPG: Fast clustering of millions of single cells. *bioRxiv* (2020).
8. Cakir, B. *et al.* Comparison of visualization tools for single-cell RNAseq data. *NAR Genomics Bioinforma.* **2,** (2020).
9. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32,** 381–386 (2014).
10. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560,** 494–498 (2018).
11. Lun, A. T. L., Richard, A. C. & Marioni, J. C. Testing for differential abundance in mass cytometry data. *Nat. Methods* **14,** 707–709 (2017).
12. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn.* **9,** 2579–2605 (2008).
13. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3,** 861 (2018).
14. Aliverti, E. *et al.* Projected t-SNE for batch correction. *Bioinformatics* **36,** 3522–3527 (2020).
15. Regev, A. *et al.* The human cell atlas. *Elife* **6,** (2017).
16. Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562,** 367–372 (2018).
17. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172,** 1091–1107.e17 (2018).
18. Wang, S. *et al.* Unifying single-cell annotations based on the Cell Ontology. *bioRxiv* (2019). doi:10.1101/810234
19. Michielsen, L., Reinders, M. & Mahfouz, A. Hierarchical progressive learning of cell identities in single-cell data. *bioRxiv* 1–9 (2020). doi:10.1101/2020.03.27.010124
20. Kiselev, V. Y., Yiu, A. & Hemberg, M. Scmap: Projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15,** 359–362 (2018).
21. Lotfollahi, M. *et al.* Query to reference single-cell integration with transfer learning. *bioRxiv* 1–26 (2020). doi:10.1101/2020.07.16.205997
22. Mourragui, S., Loog, M., Van De Wiel, M. A., Reinders, M. J. T. & Wessels, L. F. A. PRECISE: A domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. in *Bioinformatics* **35,** i510–i519 (2019).
23. Cao, K., Bai, X., Hong, Y. & Wan, L. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics* **36,** i48–i56 (2020).
24. Singh, R. *et al.* Unsupervised manifold alignment for single-cell multi-omics data. *bioRxiv* 1–14 (2020). doi:10.1101/2020.06.13.149195
25. Stark, S. G. *et al.* SCIM: Universal Single-Cell Matching with Unpaired Feature Sets. *bioRxiv* (2020).
26. Wang, W. *et al.* Live-cell imaging and analysis reveal cell phenotypic transition dynamics inherently missing in snapshot data. *Sci. Adv.* **6,** (2020).

27. Efremova, M. & Teichmann, S. A. Computational methods for single-cell omics across modalities. *Nature Methods* **17,** 14–17 (2020).
28. Method of the Year 2019: Single-cell multimodal omics. *Nature methods* **17,** 1 (2020).
29. Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nature Methods* **17,** 11–14 (2020).
30. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14,** 865–868 (2017).
31. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science (80-.).* **361,** 1380–1385 (2018).
32. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37,** 1452–1457 (2019).
33. Clark, S. J. *et al.* ScNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells e. *Nat. Commun.* **9,** (2018).
34. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nat. Methods* **14,** 414–416 (2017).
35. Kim, N. *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat. Commun.* **11,** (2020).
36. Liu, X. *et al.* A comparison framework and guideline of clustering methods for mass cytometry data. *Genome Biol.* **20,** (2019).
37. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21,** 12 (2020).
38. Hou, W., Ji, Z., Ji, H. & Hicks, S. C. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol.* **21,** (2020).
39. Chen, H. *et al.* Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.* **20,** (2019).

SUMMARY

Single-cell technologies have emerged as powerful tools to analyze complex tissues at the single-cell resolution, resolving the cellular heterogeneity within a tissue through the discovery of different cell populations. Over the past decade, single-cell technologies have greatly developed allowing the profiling of various molecular features including genomics, transcriptomics and proteomics. These high-throughput technologies produce datasets containing thousands to millions of cells in a single experiment. These large high-dimensional datasets impose several challenges to the data analysis. These challenges can be divided into three categories: **interaction**, **identification** and **integration**. **Interaction** refers to the visual exploration and interactive analysis of the data, **identification** refers to the definition of the identity of each single-cell, while **integration** deals with the combination of different molecular information from different datasets.

In this thesis, we introduced several computational methods, addressing these three challenges, to eventually improve the analysis of single-cell data. Regarding the **interaction**, we focused on developing scalable methods that can analyze datasets having millions of cells and thousands of features within workable time frames. We improved the scalability of both clustering and visualization of single-cell data by summarizing the data using a hierarchical representation.

To improve the **identification** of cells, we make use of the large number of annotated datasets available nowadays, and identify cell populations present in a single-cell dataset using classification methods instead of clustering the data. These classification methods can be trained using the previously annotated datasets. We benchmarked a large number of different classification methods and based on this analysis propose to use simple linear classifiers since they have better performance and scale better to larger datasets. We applied this linear classification on single-cell mass cytometry data to automatically identify cell populations when comparing two cohorts of colorectal cancer patients.

To **integrate** single-cell multi-omics data, we focused on extending the number of measured features to overcome current technological limitations. For single-cell mass cytometry, we integrated different panels measured from the same biological sample, resulting in an extended number of proteins markers per cell. Downstream analysis of this data revealed new cell subpopulations showing a more fine-grained cellular heterogeneity. We also extended spatial single-cell transcriptomic data by integrating it with scRNA-seq data that lacks the spatial localization of the cells. Our proposed integration generates whole transcriptome spatial data, which makes it possible to predict spatial expression patterns of genes (in-silico) that are not originally measured in the spatial data.

Taken together, this thesis presents several computational methods that aid and improve single-cell data analysis, increasing our insights in molecular heterogeneity.

SAMENVATTING

Eencellige meettechnologieën zijn krachtige geavanceerde meetinstrumenten geworden voor het analyseren van complexe weefsels waarbij inzichten in de heterogeniteit van weefsels worden vergroot door de ontdekking van nieuwe en verschillende populaties van cellen. In het afgelopen decennium hebben deze eencellige meettechnologieën zich sterk ontwikkeld waardoor op grote schaal een verscheidenheid aan moleculaire kenmerken gemeten kan worden, waaronder metingen aan het DNA en aanwezige transcripten en eiwitten. Deze geavanceerde meettechnologieën produceren datasets met meetgegevens over duizenden cellen in één experiment waarbij er per cel duizenden kenmerken gemeten zijn. Deze grote en hoog-dimensionale datasets vormen een grote uitdaging wanneer deze gegevens geanalyseerd moeten worden. Hierbij maken wij onderscheid tussen de interactie, identificatie en integratie van deze datasets. Interactie verwijst naar de visuele verkenning en interactieve analyse van de gegevens. Identificatie naar het definiëren van de identiteit van elke unieke cel. En integratie refereert naar de combinatie van verschillende moleculaire informatie uit verschillende datasets. In dit proefschrift introduceren we verschillende methoden die deze drie uitdagingen aanpakken.

Wat betreft interactie, hebben we ons gericht op het ontwikkelen van schaalbare methoden die in staat zijn om datasets met miljoenen cellen binnen een redelijk tijdsbestek te analyseren. Hierbij hebben we de schaalbaarheid van gegevensgroepering en visualisatie verbeterd door gegevens samen te vatten met behulp van een hiërarchische representatie van de data.

Om de identificatie van cellen te verbeteren, maken we gebruik van het grote aantal geannoteerde datasets dat tegenwoordig beschikbaar is. Met behulp van een classificatiemethode kunnen we dan de aanwezige celpopulaties identificeren. Deze classificatiemethoden kunnen we trainen met behulp van de eerder geannoteerde datasets. Wij hebben een groot aantal verschillende classificatiemethoden vergeleken en op basis van deze analyse komen wij tot de conclusie dat een eenvoudige lineaire classificatiemethode betere prestatie geeft en beter geschaald kan worden naar grotere datasets. We hebben deze lineaire classificatie toegepast op eencellige massa-cytometrie metingen om cellen automatisch te identificeren bij het vergelijken van twee cohorten van colorectale kankerpatiënten.

Om eencellige gegevens te integreren hebben we ons gericht op het uitbreiden van het aantal gemeten kenmerken voor die meettechnologieën waarvoor het aantal metingen per cel nog een beperkingen is. Zo hebben we voor eencellige massa-cytometrie metingen - waarbij een beperkt aantal eiwitten per cel gemeten kunnen worden - de data geïntegreerd van verschillende metingen aan hetzelfde biologisch monster waarbij iedere keer een andere collectie van eiwitten gemeten wordt. Door onze voorgestelde integratie weten we dan een groter aantal eiwitmarkers per cel. Met de analyse van deze geïntegreerde dataset hebben we nieuwe populaties van cellen gevonden die een - tot dan toe - meer fijnmazige cellulaire heterogeniteit van het monster aantoonde. Daarnaast hebben we ook eencellige ruimtelijke transcript gegevens uitgebreid – waarbij een beperkt aantal transcripten tegelijk gemeten worden - door deze te integreren met eencellige transcript gegevens die de ruimtelijke lokalisatie van de cellen missen maar wel het volledige transcriptoom van een cel meten. Onze voorgestelde integratie genereert uiteindelijk een ruimtelijke patroon van alle

transcripten waardoor patronen kunnen worden voorspeld die oorspronkelijk niet in ruimtelijke gegevens werden gemeten.

Samengevat presenteert dit proefschrift verschillende computationele methoden die de analyse van eencellige data ondersteunen en verbeteren waardoor ons inzicht in moleculaire heterogeniteit wordt vergroot.

ACKNOWLEDGMENTS

الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ

Over the past 4 years, I had the opportunity to enrich my knowledge and proceed in my career by pursuing my PhD, which is something I was looking forward to achieve long time ago. As a person who enjoy learning, I enjoyed most of my PhD journey as for me it was a continuous learning curve. But it was indeed a tough journey, and a lot different than I thought. As much setbacks and hard times, I had the privilege to travel around the world to beautiful places and communicate with wonderful people, each single person added a unique piece to my knowledge and/or my personality which was essential to make my journey as successful as it was.

First and foremost, I would like to thank my family for their unlimited efforts and support. I would have never achieved anything without my parents, **Roshdy** and **Souzan**. Thank you for everything, your constant encouragement and support made me what I am today. I hope I made you proud, that is the least I can offer you.

أبي رشدى وأمي سوزان، لم أكن لأحقق أي شيء بدونكما، شكراً لكما على كل شيء، إن تشجيعكما المستمر ودعمكما جعلني ما أنا عليه اليوم. أتمنى أن يكون كلاكما فخوراً بي، هذا أقل ما يمكنني تقديمه لكما.

My sisters **Dalia** and **Ingy**, you will always be my best friends and I will keep bothering you. Special thanks to **Dalia** for all the support she gave me when I first arrived to the Netherlands.

To my beloved little family: my beautiful wife and lifetime partner **Donia**, thank you for all your help, support and patience. I could have never done it without you, I cannot express how grateful I am for all your love, care and sacrifices. My kids **Fares** and **Fayrouz**, you are the most beautiful gift I have in my life, God bless you and I hope one day you will be proud of your father.

Regardless of all the hard work that is done, most of the credit belongs to my supervisors **Marcel** and **Ahmed**. I want to thank you for all your guidance, teaching, support and professional/personal advices. I still remember when I first joined the group from a relatively different field with almost zero background in bioinformatics. Thanks to you I was able to slowly grasp everything and successfully finish my PhD. **Marcel**, I really admire your critical way of thinking and your challenging discussions with me, you were always pushing me further, thank you for being such a great mentor. **Ahmed**, working with you made me always confident to never get stuck, I always knew that I can find simple solutions when discussing my problems with you, I'm grateful for everything I've learned from you.

Boudewijn, my unofficial supervisor, thank you for all the guidance and collaborations we've had together, happy to continue working with you and the rest of the group. **Thomas**, teaching with you was quite an experience that I will never forget, specially how to deal with students. **Joana**, I enjoyed our endless discussions at your office door. **Thomas Höllt**, thank you for all the help with my continuous visualization related questions. **Erik**, thank you for the useful discussions we always have at the coffee machine.

Many thanks for my friends and colleagues in the PRB group. **Arlin**, or maybe I should say Guitarlin, sharing the office with you from day one was really a lot of fun, good luck with your current job and with your music. **Stavros**, my buddy, you are a true friend and a great colleague. I learn a lot from our non/scientific discussions, and I really enjoyed your company in and outside the office, playing basketball together and traveling around. **Ramin**, the stairs climber with great Persian cooking and goalkeeping skills, being in the same office was quite fun, I really have a great time with you. **Mostafa**, a true lifetime friend, when you joined the group I was probably more happy than you did, I hope you don't hate me already for this, and for another reason. I really cannot express how much positive impact your move here had on me. **Soufiane**, thank you for the great collaboration we had in the SpaGE paper. I've learned a lot from you and from the interesting discussions we've had together, also playing football with you was quite fun. **Lieke**, working with you on the benchmarking study was probably one of the best periods in my PhD, I really like our continuous discussions about each other projects and the ideas we always share. I hope we can do something together again in the future.

Alex, a.k.a Sally, the one person who almost know how to play all kind of sports, I really enjoyed playing basketball and football with you, please take care of your shoulder. **Christian**, my Friday afternoon companion in the old EWI when everyone is already home, you are almost the only person who managed to escape from my continuous tries to make you join the football, based on the injury records I think you did the right thing. **Christine**, loudest laugh ever, I could almost know if you are in the office or not without bothering to leave my chair, I really admire the social positive energy you're adding to the group. **Tom**, thanks for inviting me to try bouldering, it was quite fun. We've also had a lot of laughs about so many ridiculous papers. **Mohammed**, when you joined the group as a master student you were extremely quiet and mysterious, it turns out later that you're a great guy. Thank you for organizing the reading group but please next time don't pick books that makes us suffer. **Amelia**, you visited us for a short time but you indeed had a great impact, thank you for all the good memories. **Aysun**, I enjoyed all the funny and serious discussions we had over borrels, and indeed the church bells. **Nicco**, the highly skilled Italian football player, in our first retreat we got assigned to work together on a topic which I had no idea about, I think we did good. **Lucas**, thank you for introducing me to Explosive kittens, I really liked that game. Thanks to the DBL-Amsterdam people **Jasper**, **Marc**, **Sven**, **Meng** and **Henne** for all the useful discussions. **Sjoerd**, although we only shared office for few months, we had some nice discussions and I bothered you with many questions about how to use the cluster. **Thies**, on my first week in TUDelft you gave me a valuable advice saying that if you have three months left in your contract and your supervisors asked you to start a new project say NO, I did make use of that towards the end of my PhD. **Amin**, thank you for showing me the prayer room in the old EWI. Special thank you to **Saskia**, **Bart**, **Ruud** and **Robbert** for the great support they offer to the whole group to make sure everything goes well.

Osman, the back heel football guy, you were a valuable member of the PRB sports Committee, thanks for all the nice time we've had together in the office and on the football pitch. **Yancong**, my floor mate during lockdown time and by far PRB best football defender, moving to the same office was a turning point in knowing you for real, thanks for all the fun time we had together. **Ekin**, I still don't understand how you managed to break your finger in a slow-motion goalkeeping move. **Laura**, you were always keen about the well-being of the group members, every Wednesday you would ask me if everyone is still alive after Tuesday-football. **Stephanie**, thank you for providing us with your PS controller for the DBL retreat. **Robert-Jan**, I really enjoy our chat during borrels, especially your explanation of any topic related to the Dutch government. **Jose**, I enjoyed playing football with you, and the nice discussions we had during the poster sessions. **Chirag**, you have great social skills, however,

your football skills are not on the same level, especially scoring in an empty net. I really had great time with you on the pitch, hope we do it again soon. **Taygun**, playing basketball with you was quite fun. Giving your football skill level, I'm still surprised you didn't know who is Mo Salah, yes he is not my brother. I would also like to thank everyone from the PR/CV/SPC labs for the nice discussions we had during the poster sessions, the few coffee talks that I was able to join, and the great time during the Thursday borrels: **David**, **Marco**, **Jan**, **Hayley**, **Wouter**, **Wenjie**, **Yanxia**, **Yazhou**, **Seyran**, **Bob**, **Hamdi**, **Amogh**, **Silvia**, **Jesse**, **Xin**, PR **Tom**, **Alexander**, **Arman**, **Jin**, **Yunqiang**, **Ombretta**, **Nergis**, **Attila**, **Yeshwanth**, **Ziqi**, **Burak** and **Marian**.

For the new generation of the PRB who joined the group during the lockdown, **Yasin** (the skilled basketball player), **Colm**, **Attila Csala**, **Gerard**, PR **Ramin**, DBL **Stephanie** and **Skander**, we already started to know each other but mostly virtual. I hope in the near future we can see each other in person and get to know each other better.

Further, I want to thank all the members of the LCBC and MOLEPI groups in LUMC for all the interesting discussions and the fun time we had: **Indu**, **Mikhael**, **Laura Heezen**, **Dongxu**, **Daniele**, **Martijn**, **Antonis**, **Fatih**, **Janine**, **Rodrigo**, **Leon** and **Davy**. Special thank you to **Paul de Raadt**, who owns most of the credit for chapter 2 of this book. I would like to thank all the great immunologists in LUMC that I had the chance to collaborate with them: **Vincent**, for the great guidance in basics of immunology, **Natasja**, for the continuous collaboration and exchange of knowledge we do, also for the fun time we had together in Maastricht and Stockholm, **Li Na**, **Guillaume**, **Esmé**, for being part of your interesting studies, **Frits**, **Noel**, and **Ramon**. During my visit to Bergen, I had to chance to meet and collaborate with wonderful scientists, I would like to thank you all for the hospitality and the great collaboration we had together: **Sam**, **Liv Cecilie**, **Stein-Erik**, **Katrin**, **Emmet** and **Line**.

To my friends in Egypt, **Islam Mohsen**, **Amir**, **Mostafa**, **Medhat**, **Ehab** and **Islam Youssef**, thank you for all the support and the fun time we spend together, which always heals me and fills me with positive energy to keep going forward. Also, I cannot forget to mention **Walid**, **Kilany** and **Bially**, thank you for all the lovely gatherings and outings we had together in the Netherlands, which made me feel like home.

Finally, I want to thank all my professors and teachers from the faculty of Engineering in Cairo University for everything they taught me. Special thank you to **Dr. Sherif Sami** and **Dr. Ayman Eldeib** for being great mentors and supervisors during my masters.

CURRICULUM VITÆ

Tamim Abdelaal was born June 2nd 1989 in Cairo, Egypt. In 2011, Tamim finished his Bachelor's degree and graduated from Biomedical Engineering and Systems department, Cairo University, Egypt. Afterwards, Tamim worked as Teaching Assistant and started his Masters studies in the same department, where his research was in the field of Bioinstrumentation. As a master student, Tamim joined the Medical Equipment Calibration Laboratory where he worked as testing and calibration engineer, and became the Technical Manager of the Laboratory two years later.

After obtaining his Master's degree in 2015, Tamim started his PhD studies at the Delft Bioinformatics Lab, Faculty of Electrical Engineering Mathematics and Computer Science at the TU Delft, the Netherlands. His research was focused on developing computational methods to improve the data analysis of various single-cell technologies such as mass cytometry, scRNA-seq and spatial transcriptomics. During his PhD, Tamim worked jointly between the Delft Bioinformatics Lab and the Leiden Computation Biology Center (LCBC) at the Leiden University Medical Center (LUMC), and had many collaborations with the Department of Immunohematology and Blood Transfusion at the LUMC. In 2018, Tamim visited the Department of Clinical Science at the University of Bergen, Norway. During his two months visit, Tamim worked on analyzing different single-cell dissociation methods for mass cytometry panel designed to study ovarian cancer.

Starting October 2020, Tamim works as a post-doctoral researcher at the Department of Radiology at the LUMC, the Netherlands. Tamim currently works on methods to study cellular differentiation in the spatial context of the tissue, as well as comparative analysis of different spatial transcriptomics protocols studying their ability to accurately map the spatial distributions of cortical cell types in the mouse brain.

PUBLICATIONS

T. Abdelaal, P. de Raadt, B.P.F. Lelieveldt, M.J.T. Reinders, A. Mahfouz, "SCHNEL: scalable clustering of high dimensional single-cell data", *Bioinformatics* (2020)

T. Abdelaal, J. Eggermont, T. Höllt, A Mahfouz, M.J.T. Reinders, B.P.F. Lelieveldt, "Cytosplore-Transcriptomics: a scalable inter-active framework for single-cell RNA sequencing data analysis", *Biorxiv* (2020)

E.T.I. van der Gracht, G. Beyrend, **T. Abdelaal**, I.N. Pardieck, T.H. Wesselink, F.J. van Haften, S. van Duikeren, F. Koning, R. Arens, "Memory CD8+ T cell heterogeneity is primarily driven by pathogen-specific cues and additionally shaped by the tissue environment", *iScience* (2020)

T. Abdelaal, S. Mourragui, A. Mahfouz, M.J.T. Reinders, "SpaGE: Spatial Gene Enhancement using scRNA-seq", *Nucleic Acid Research* (2020)

T. Abdelaal, L. Michielsen, D. Cats, D. Hoogduin, H. Mei, M.J.T. Reinders, A. Mahfouz, "A comparison of automatic cell identification methods for single-cell RNA sequencing data", *Genome Biology* (2019)

N. Li, V. van Unen, N. Guo, **T. Abdelaal**, A. Somarakis, J. Eggermont, A. Mahfouz, S.M. Chuva de Sousa Lopes, B.P.F. Lelieveldt, F. Koning, "Early-Life Compartmentalization of Immune Cells in Human Fetal Tissues Revealed by High-Dimensional Mass Cytometry", *Frontiers in Immunology* (2019)

N.L. de Vries, V. van Unen, M.E. Ijsselsteijn, **T. Abdelaal**, R. van der Breggen, A.F. Sarasqueta, A. Mahfouz, K.C.M.J. Peeters, T. Höllt, B.P.F. Lelieveldt, F. Koning, N.F.C.C. de Miranda, "High-dimensional cytometric analysis of colorectal cancer reveals novel mediators of antitumour immunity", *Gut* (2019)

T. Abdelaal, T. Höllt, V. van Unen, B.P.F. Lelieveldt, F. Koning, M.J.T. Reinders, A. Mahfouz, "CyTOFmerge: integrating mass cytometry data across multiple panels", *Bioinformatics* (2019)

T. Abdelaal, V. van Unen, T. Höllt, F. Koning, M.J.T. Reinders, A. Mahfouz, "Predicting cell populations in single cell mass cytometry data", *Cytometry Part A* (2019)

N. Li, V. van Unen, **T. Abdelaal**, N. Guo, S.A. Kasatskaya, K. Ladell, J.E. McLaren, E.S. Egorov, M. Izraelson, S.M. Chuva de Sousa Lopes, T. Höllt, O.V. Britanova, J. Eggermont, N.F.C.C. de Miranda, D.M. Chudakov, D.A. Price, B.P.F. Lelieveldt, F. Koning, "Memory CD4+ T cells are generated in the human fetal intestine", Nature Immunology (2019)