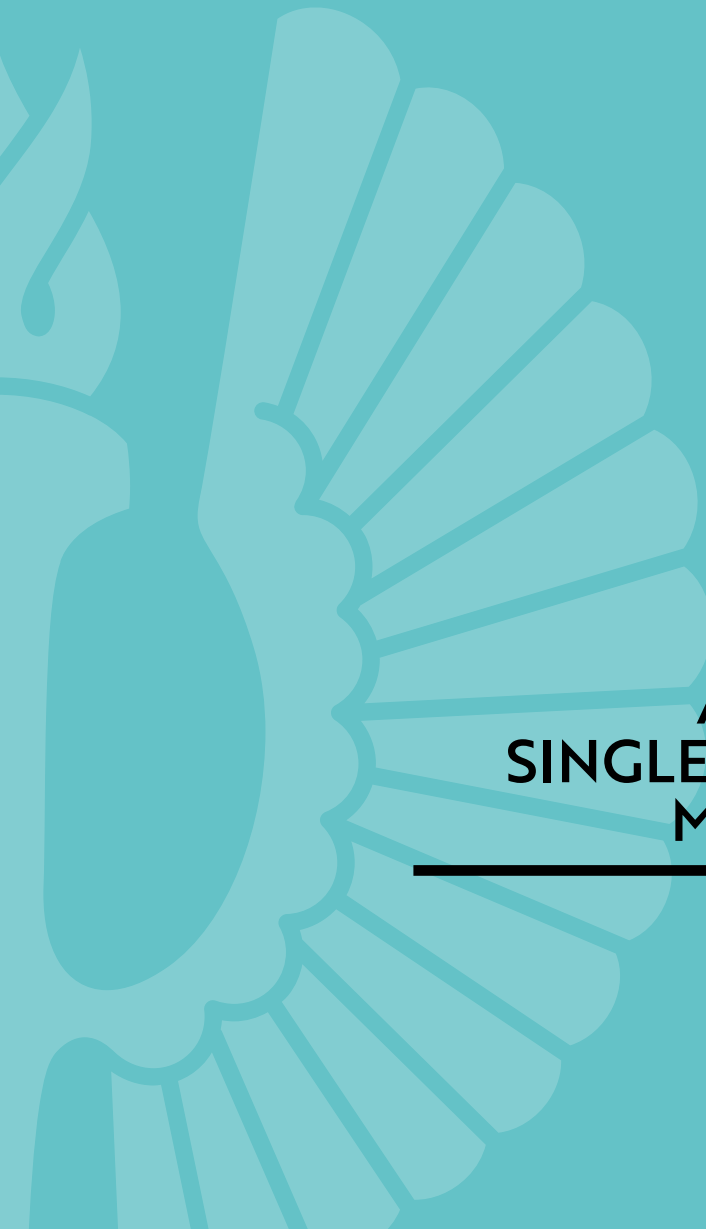




**TURUN
YLIOPISTO**
UNIVERSITY
OF TURKU

A large, light blue, stylized sunburst or fan-like graphic is positioned on the left side of the page. It consists of a central vertical stem with multiple curved, radiating segments extending outwards, creating a fan-like appearance.

COMPUTATIONAL APPROACHES FOR SINGLE-CELL OMICS AND MULTI-OMICS DATA

Nigatu Ayele Adossa



**TURUN
YLIOPISTO**
UNIVERSITY
OF TURKU

COMPUTATIONAL APPROACHES FOR SINGLE-CELL OMICS AND MULTI-OMICS DATA

Nigatu Ayele Adossa

University of Turku

Faculty of Technology
Department of Computing
Computer Science
Doctoral Programme in Technology

Supervised by

Professor, Laura Elo
Turku Bioscience Centre
University of Turku and
Åbo Akademi University
Turku, Finland

Docent, Dr, Kalle Rytönen
Turku Bioscience Centre
University of Turku
Turku, Finland

Reviewed by

Docent, Reija Autio
Tampere University
Tampere, Finland

Dr, Johan Henriksson
Umeå University
Umeå, Sweden

Opponent

Professor, Sascha Ott
University of Warwick
Coventry, United Kingdom

The originality of this publication has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

ISBN 978-951-29-9449-6 (PRINT)
ISBN 978-951-29-9450-2 (PDF)
ISSN 2736-9390 (Painettu/Print)
ISSN 2736-9684 (Sähköinen/Online)
Painosalama, Turku, Finland 2023

Desalech Dana (1967–2000)

UNIVERSITY OF TURKU

Faculty of Technology

Department of Computing

Computer Science

NIGATU AYELE ADOSSA: Computational Approaches for Single-cell Omics and Multi-omics Data

Doctoral Dissertation, 133 pp.

Doctor of Philosophy

April 2023

ABSTRACT

Single-cell omics and multi-omics technologies have enabled the study of cellular heterogeneity with unprecedented resolution and the discovery of new cell types. The core of identifying heterogeneous cell types, both existing and novel ones, relies on efficient computational approaches, including especially cluster analysis. Additionally, gene regulatory network analysis and various integrative approaches are needed to combine data across studies and different multi-omics layers. This thesis comprehensively compared Bayesian clustering models for single-cell RNA-sequencing (scRNA-seq) data and selected integrative approaches were used to study the cell-type specific gene regulation of uterus. Additionally, single-cell multi-omics data integration approaches for cell heterogeneity analysis were investigated.

Article I investigated analytical approaches for cluster analysis in scRNA-seq data, particularly, latent Dirichlet allocation (LDA) and hierarchical Dirichlet process (HDP) models. The comparison of LDA and HDP together with the existing state-of-art methods revealed that topic modeling-based models can be useful in scRNA-seq cluster analysis. Evaluation of the cluster qualities for LDA and HDP with intrinsic and extrinsic cluster quality metrics indicated that the clustering performance of these methods is dataset dependent.

Article II and **Article III** focused on cell-type specific integrative analysis of uterine or decidual stromal (dS) and natural killer (dNK) cells that are important for successful pregnancy. **Article II** integrated the existing preeclampsia RNA-seq studies of the decidua together with recent scRNA-seq datasets in order to investigate cell-type-specific contributions of early onset preeclampsia (EOP) and late onset preeclampsia (LOP). It was discovered that the dS marker genes were enriched for LOP downregulated genes and the dNK marker genes were enriched for upregulated EOP genes. **Article III** presented a gene regulatory network analysis for the subpopulations of dS and dNK cells. This study identified novel subpopulation specific transcription factors that promote decidualization of stromal cells and *dNK* mediated maternal immunotolerance.

In **Article IV**, different strategies and methodological frameworks for data integration in single-cell multi-omics data analysis were reviewed in detail. Data integration methods were grouped into early, late and intermediate data integration strategies. The specific stage and order of data integration can have substantial effect on the results of the integrative analysis. The central details of the approaches were presented, and potential future directions were discussed.

KEYWORDS: Clustering, Single-cell RNA-sequencing, Single-cell omics, Uterus, Preeclampsia

TURUN YLIOPISTO

Tietojenkäsittelytieteen laitos

Teknillinen Tiedekunta

Tietokone Tiede

NIGATU AYELE ADOSSA: Laskennallisia menetelmiä yksisolusekvensointi- ja multiomiikkatulosten analyysiin

Väitöskirja, 133 s.

Filosofian Tohtori

Huhtikuu 2023

TIIVISTELMÄ

Yksisolusekvensointitekniikat mahdollistavat solujen heterogeenisyyden tutkimuksen ennennäkemättömällä resoluutiolla ja uusien solutyypin löytämisen. Solutyypin tunnistamisessa keskeisessä roolissa on ryhmittely eli klusterointianalyysi. Myös geenien säätelyverkostojen sekä eri molekyyliatasojen yhdistäminen on keskeistä analyysissä. Väitöskirjassa verrataan bayesilaisia klusterointimenetelmiä ja yhdistetään eri menetelmillä kerättyjä tietoja kohdan solutyypispesifissä geeninsäätelyanalyysissä. Lisäksi yksisolutiedon integraatiomenetelmiä selvitetään kattavasti.

Julkaisu I keskittyy analyttisten menetelmien, erityisesti latenttiin Dirichlet-allokaatioon (LDA) ja hierarkkiseen Dirichlet-prosessiin (HDP) perustuvien mallien tutkimiseen yksisoludatan klusterianalyysissä. Kattava vertailu näiden kahden mallin sekä olemassa olevien menetelmien kanssa paljasti, että aiemallinnuspohjaiset menetelmät voivat olla hyödyllisiä yksisoludatan klusterianalyysissä. Menetelmien suorituskyky riippui myös kunkin analysoidun datasetin ominaisuuksista.

Julkaisuissa II ja III keskitytään naisen lisääntymisterveydelle tärkeiden kohdan stroomasolujen ja NK-immunisolujen solutyypispesifiseen analyysiin. Artikkelissa II yhdistettiin olemassa olevia tuloksia pre-eklampsiasta viimeisimpiin yksisolusekvensointituloksiin ja löydettiin varhain alkavan pre-eklampsian (EOP) ja myöhään alkavan pre-eklampsian (LOP) solutyypispesifisiä vaikutuksia. Havaittiin, että erilaistuneen strooman markkerigeenien ilmentyminen vähentyi LOP:ssa ja NK-markkerigeenien ilmentyminen lisääntyi EOP:ssa. **Julkaisu III** analysoi strooman ja NK-solujen alapopulaatiospesifisiä geeninsäätelyverkostoja ja niiden transkriptiofaktoreita. Tutkimus tunnisti uusia alapopulaatiospesifisiä säätelijöitä, jotka edistävät strooman erilaistumista ja NK-soluvälitteistä immunotoleranssia

Julkaisu IV tarkastelee yksityiskohtaisesti strategioita ja menetelmiä erilaisten yksisoludatatasojen (multi-omiikka) integroimiseksi. Integrointimenetelmät ryhmiteltiin varhaisen, myöhäisen ja välivaiheen strategioihin ja kunkin lähestymistavan menetelmiä esiteltiin tarkemmin. Lisäksi keskusteltiin mahdollisista tulevaisuuden suunnista.

ASIASANAT: klusterointi, yksisolu-RNA-sekvensointi, yksisolu-omiikka, kohtu, pre-eklampsia

4.1	Datasets.....	36
4.1.1	Human Artificially Mixed Immune Dataset	36
4.1.2	Mouse Cell Atlas Dataset	36
4.1.3	Human 1 st Trimester Pregnancy Data.....	36
4.1.4	Human Menstrual Cycle Data.....	37
4.1.5	Human Term Pregnancy Data.....	37
4.1.6	Pregnancy disorder datasets.....	37
4.1.7	Human cisTarget Motifs.....	37
4.2	Methods and Analytical Workflow.....	38
4.2.1	Comparison of LDA and HDP Models for Single-cell RNA-seq Clustering (Article I).....	38
4.2.1.1	Measures of Cluster Quality.....	39
4.2.2	scRNA-seq Cluster Analysis of Decidual Cells (Article II & Article III).....	41
4.2.3	Gene Over-representation Analysis (Article II and Article III).....	42
4.2.4	Single-cell Gene Regulatory Network Analysis (Article III).....	42
4.2.5	The Review of Single-cell Multi-omics Data Integration (Article IV).....	43
5	Results.....	44
5.1	Comparison of LDA and HDP for scRNA-seq Clustering	44
5.1.1	Clustering Performance.....	44
5.1.2	Computational Scalability	46
5.1.3	Comparison of LDA Clustering Tools.....	46
5.2	Contributions of Cell-type Specific Markers in Preeclampsia	47
5.2.1	Contribution of Decidual Stromal Cell Subpopulation Markers in Preeclampsia	47
5.2.2	Contributions of Decidual Natural Killer Cell Subpopulation Markers in Preeclampsia	49
5.3	Decidualization Regulatory Network Inference for Stromal and Natural Killer Cells.....	50
5.3.1	Subpopulations of Stromal and NK Cells.....	50
5.3.2	Regulatory Networks of Decidualizing Stromal Cells ...	51
5.3.3	Regulatory Networks for Decidual NK Cells.....	53
5.3.4	Decidual Stromal and NK Cell Regulators in Pregnancy Disorders.....	55
5.4	Single-cell Multi-omics Integrative Data Analysis.....	55
5.4.1	Single-cell Multi-omics Data Integration Strategies.....	55
5.4.2	Early Data Integration Approaches.....	56
5.4.3	Late Data Integration Approaches.....	56
5.4.4	Intermediate Integration Methods.....	57
5.4.4.1	Similarity-based Methods.....	57
5.4.4.2	Joint Dimensionality Reduction.....	57
5.4.4.3	Model-based Methods.....	58
5.5	Contributions of the thesis.....	58
6	Discussion.....	60
6.1	Limitations.....	62

6.2 Future directions	63
7 Conclusions	64
Acknowledgements.....	65
List of References	66
Original Publications.....	81

Abbreviations

ARI	Adjusted Rand Index
AUC	Area Under the Curve
CCA	Canonical Correlation Analysis
CH-index	Calinski–Harabasz Index
DB-index	Davies-Bouldin Index
dNK	Decidual Natural Killer cell
DP	Dirichlet Process
DPMM	Dirichlet Process Mixture Model
dS	Decidual Stromal Cell
EOP	Early Onset Preeclampsia
GRN	Gene Regulatory Network
HDP	Hierarchical Dirichlet process
LDA	Latent Dirichlet Allocation
LOP	Late Onset Preeclampsia
M-PP	Menstrual cycle from women Previously had Preeclampsia
MCMC	Markov Chain Monte Carlo
MDS	Multidimensional Scaling
NGS	Next Generation Sequencing
NMF	Non-Negative Matrix Factorization
PCA	Principal Component Analysis
QC	Quality Control
RSS	Regulon Specificity Score
scATAC-seq	Single-cell ATAC Sequencing
scMulti-omics	Single cell Multi-omics
scRNA-seq	Single cell RNA Sequencing
TF	Transcription Factor
SNN	Shared Nearest Neighbor
UMAP	Uniform Manifold Approximation and Projection for Dimension Reduction
UMI	Unique Molecular Identifiers
WGA	Whole Genome Amplification

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

- I. Adossa N, Rytönen KT, Elo LL. Dirichlet process mixture models for single-cell RNA-seq clustering. *Biol Open*. 2022;11(4): bio059001.doi:10.1242/BIO.059001
- II. Rytönen KT*, Adossa N*, Mahmoudian M, Lönnberg T, Poutanen M, Elo LL. Cell type markers indicate distinct contributions of decidual stromal cells and natural killer cells in preeclampsia. *Reproduction*. 2022;164(5):V9-V13. doi:10.1530/REP-22-0079
- III. Rytönen KT*, Adossa N*, Zúñiga Norman S*, Lönnberg T, Poutanen M, Elo LL. Gene regulatory network analysis of decidual stromal cells and natural killer cells. Manuscript.
- IV. Adossa N, Khan S, Rytönen KT, Elo LL. Computational strategies for single-cell multi-omics integration. *Comput Struct Biotechnol J*. 2021;19:2588–2596. Doi:10.1016/j.csbj.2021.04.060

The original publications have been reproduced with the permission of the copyright holders. * Equal first author contribution

1 Introduction

Single-cell sequencing technologies have recently embarked on a new scientific arena in the biological, biomedical, and medical research communities with their wide varieties of applications in developmental biology, cancer biology, immunology, microbial research, etc. Since the emergence of sanger sequencing in the early 1970s, Next Generation Sequencing (NGS) technology has advanced tremendously. Currently, it is possible to do bulk whole genome sequencing for a cheaper price within a short period of time enabling wider accessibility of such technologies for medical and biological research. Moreover, the recent single-cell technological advances, harnessing microfluidics and other cell sorting technologies, have made a significant contribution towards sequencing massive numbers of cells across samples at a single-cell resolution.

The pitfall of bulk sequencing technologies comes from its potential to hinder the contributions of individual cells by quantifying sample-level averages. Therefore, it is challenging to study the heterogeneity among cells in the sample group or tissue sample. However, the single-cell sequencing technologies avoid average quantification and quantify the omics measurement at single-cell resolution. This allows the study and identification of subpopulations and/or cell states. In addition, it allows researchers to uncover new and potentially unexpected biological discoveries, such as revealing complex and rare cell populations, identifying regulatory relationships between genes, and tracking the trajectories of distinct cell lineages in development that the traditional bulk sequencing fails.

Additionally, the application of single-cell technologies toward elucidating the cellular heterogeneity from different omics layers such as genomics, epigenomics, transcriptomics, and proteomics has tremendous potential for unlocking the unknowns in biomedical, pharmaceutical and medical research areas. In developmental biology, single-cell RNA-seq (scRNA-seq) [1–8] has been used to get insights into early embryonic development. scRNA-seq is also a widely used protocol in the study of tumor heterogeneity study in cancer research [9–15], immunology [16–22] and evolutionary lineage tracing [23,24]. Single-cell epigenomic protocols such as single-cell ATAC sequencing (scATAC-seq) [25], and single-cell bisulfite sequencing (sc-BS-seq) [26] are also widely used in recent

years to uncover the cellular heterogeneity from the epigenomic landscape. Although single-cell genomics [27–33] and proteomics [34–36] are not widely utilized as single-cell transcriptomics [37,38], they have immense potential to uncover cellular heterogeneity from genomic and proteomic contexts.

However, the independent profiling of a single omics data from a single cell gives a single snapshot view of the given cell at a time. This only gives partial information about the cell where the biological and molecular mechanisms are intertwined with the interactions among several molecules. In this context, the advance in single-cell sequencing technology for independent omics layer ignited a curiosity to explore the potential of multi-modal molecular profiling assay at single-cell resolution. With the emerging techniques for cell isolation and disassociation, the profiling of more than one omics layer from a single cell becomes a reality. Such an effort towards the development of multi-modal omics profiling at single-cell resolution embarked on a new era of scientific exploration in the field of molecular biology and bioinformatics. Profiling multiple omics data from a single cell created an opportunity for researchers to study multi-modal molecular assays boosting the exploration and discovery of biological mechanisms and gene regulation.

With the growing amount of both single-cell omics and multi-omics data, the computational aspect of storing, preprocessing, analyzing, visualization, and interpretation of such a massive amount of data poses computational challenges. The first part of this thesis work (**Article I**) addresses the comparison between latent Dirichlet allocation (LDA) and Hierarchical Dirichlet process (HDP) models for cell heterogeneity analysis in single-cell RNA-seq data. The second research (**Article II**) combines bulk and single-cell RNA-seq data to study the cell-type-specific contributions of marker genes in the disease called preeclampsia. Thirdly, the gene regulatory networks for the decidual stromal and natural killer cell subpopulations from single-cell RNA-seq data were analyzed to explore the cell type-specific gene regulatory networks (**Article III**). Finally, the strategies and methodological approaches for single-cell multi-omics data integration were explored (**Article IV**).

2 Review of the Literature

2.1 Single-cell Technologies

2.1.1 Single-cell Omics

The single-cell sequencing emerged as a de-facto protocol for studying cellular heterogeneity in biomedical research. The workflow starts with sample extraction from tissue or biopsy. The extracted sample undergoes the cell disassociation. Then, the molecular profiling of the desired omics type followed by PCR amplification, library preparation and sequencing takes place as illustrated in **Figure 1**.

There are several cell isolation techniques including limiting dilution [39], micromanipulation [40], flow-activated cell sorting (FACS) [41], laser capture microdissection [42], and microfluidics-based methods [43]. Limiting dilution [39] utilizes pipettes to isolate individual cells by dilution with the major limitation of only attaining about one-third of the prepared wells in a well plate. Another widely used technique is micromanipulation [40] where microscope-guided capillary pipettes are used to extract cells. It has been utilized to retrieve cells from early embryos [44]. The major drawback of such a method is that it is time-consuming and has low throughput. The method that is widely used among immunologists for cell isolation and sorting is flow-activated cell sorting (FACS) [41]. It tags the cells with a fluorescent monoclonal antibody to recognize specific cell surface markers. The need for specific monoclonal antibodies and the fact that it requires large number of starting cells can be considered as a drawback of this method.

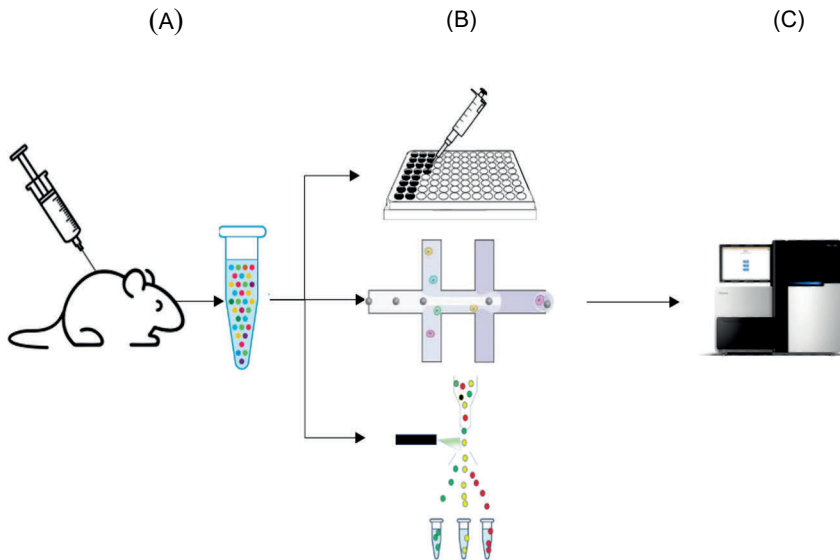


Figure 1. The single-cell omics sequencing workflow. (A) Sample extraction. (B) Cell sorting and library preparation techniques using fluorescence-activated cell sorting (FACS), microfluidics and micromanipulation techniques. (C) Finally, reads are sequenced using a sequencing machine.

Laser capture microdissection is another method for doing cell isolation from solid samples using computerized laser technologies [42]. The microfluidics-based technique [43] is the most recent and the one that has revolutionized single-cell omics sequencing by providing precise nano-liter sized fluid control for cell isolation, low sample consumption, device miniaturization, low risk of contamination, and low analysis cost. Despite all the advantages and potential of the microfluidics technique, it has a major drawback as it requires a minimum of 1000 cells to capture and has a requirement of homogeneous cell size [39].

Once the individual cell is isolated, the next step in the workflow is to profile the desired molecule from a given cell [45]. Several protocols perform molecular profiling for different omics layers i.e., transcriptomics (DroNc-seq [46], Drop-seq [47], inDrop [48], 10x Genomics [49], Nx1-seq [50] and Seq-Well [51]), genomics (MDA [52], MALBAC [53], DOP-PCR [54]), epigenomics (scBS-seq [26], scRRBS [55], ATAC-seq [56] and Hi-C [57]) and proteomics (MACS Chip [58], CyToF [59]).

2.1.1.1 Single-cell Transcriptomics Protocols

Single-cell transcriptomics is a widely used protocol to study transcriptomics-level cellular heterogeneity. There are several single-cell transcriptomics profiling techniques including plate-based, microdroplet and microwell-based protocols. The plate-based protocols such as Smart-seq1-3 [60–62] and Quartz-Seq [63] use full-length cDNA amplification with oligo-dT priming and template switching for quantification of stable mRNA molecules from an individual cell. The microdroplet and microwell-based protocols, such as DroNc-seq [46], Drop-seq [47], inDrop [48], 10x Genomics [49], Nx1-seq [50] and Seq-Well [51], are designed in such a way that a cell/nucleus barcoded bead and reaction liquid are encapsulated as oil droplets and reverse transcription take place with molecular/cell barcoding within each of the oil droplets (**Figure 1B**). The molecular barcodes or Unique Molecular Identifiers (UMIs) are used to identify the PCR duplicates computationally. Such protocols enable higher throughput by enabling the sequencing of thousands of individual cells at a relatively lower cost.

2.1.1.2 Single-cell Genomics Protocols

To profile the genomic DNA for detecting point mutations, copy number variations (CNV), and structural aberrations at single-cell resolutions, uniform whole genome amplification (WGA) techniques such as multiple displacement amplification (MDA) [52], multiple annealing and looping-based amplification cycles (MALBAC) [53] and degenerate oligonucleotide-primed PCR (DOP-PCR) [54] are widely used. However, most of such WGA methods are inefficient in achieving a uniform sequencing depth due to the amplification bias. Therefore, it is recommended to pay special attention during the data analysis [52].

2.1.1.3 Single-cell Epigenetics Protocols

There have been several protocols developed for single-cell epigenetic profiling of DNA methylation, histone modification, and chromatin accessibility. The single-cell bisulfite sequencing (scBS-seq) [26] and single-cell reduced representation bisulfite sequencing (scRRBS) [55] are protocols used to profile the whole genome and targeted DNA methylation, respectively. Drop-ChIP [64] is another recently introduced microfluidic-based epigenetic profiling protocol to investigate the chromatin state using histone modification at single-cell resolution. Assay for transposase-accessible chromatin using sequencing (ATAC-seq) [56] protocol tags an open chromatin region with a sequencing adaptor by Tn5 transposase before amplification to profile open chromatin patterns in each of the cells. Another similar

epigenetic protocol called Hi-C [57] profiles the genomics regions in a spatial proximity context in the nuclei adopting in-nucleus ligation.

2.1.1.4 Single-cell Proteomics Protocols

The single-cell profiling of proteomes has also shown significant developments in the recent past, though its throughput is limited. The single-cell protein profiling techniques can be divided into two based on the use of mass-spectrometry, ie., the antibody-based and mass-spectrometry-based techniques [65]. The antibody-based assay technique targets specific protein using tagged antibodies. The fluorescence-based assays including fluorescence flow cytometry (FFC) [66] and microfluidic antibody capture chip (MACS Chip) [58] exposes cells with different fluorescence markers that are specific to certain protein for detection and quantification of the protein of interest at a single cell level. The mass cytometry-based techniques [67] such as CyToF [59] profile single-cell intracellular and surface protein by utilizing a labeled antibody tag. Using an approach for single-cell Western blotting [68] is also among the popular antibody-based techniques for single-cell protein profiling. Such antibody-based approaches result in low-level protein multiplexing (10-15 proteins). However, the single-cell mass-spectrometry-based method such as Single Cell Proteomics by Mass Spectrometry (SCoPE-MS) [69] profiles tens to hundreds of protein expressions though there are still technical challenges with respect to the detection coverage [70].

2.1.2 Single-cell Multi-omics

Once the cells are isolated, multiple molecular extractions i.e, transcripts and genomic DNA, transcripts and epigenetic measurements, transcripts and proteins, or even more than two molecular profiling layers can be achieved from a single cell with different single-cell multi-modal protocols (**Figure 2**).

2.1.2.1 Transcriptome and Genome

One of the widely used method for simultaneous profiling of mRNA transcripts and genomic DNA is the physical separation strategy that physically isolate the nucleus from cytosolic molecules. Then, the nucleus with genomic DNA and the cytosolic component containing a significant amount of mRNA molecules are dealt with separately in the downstream protocols [71].

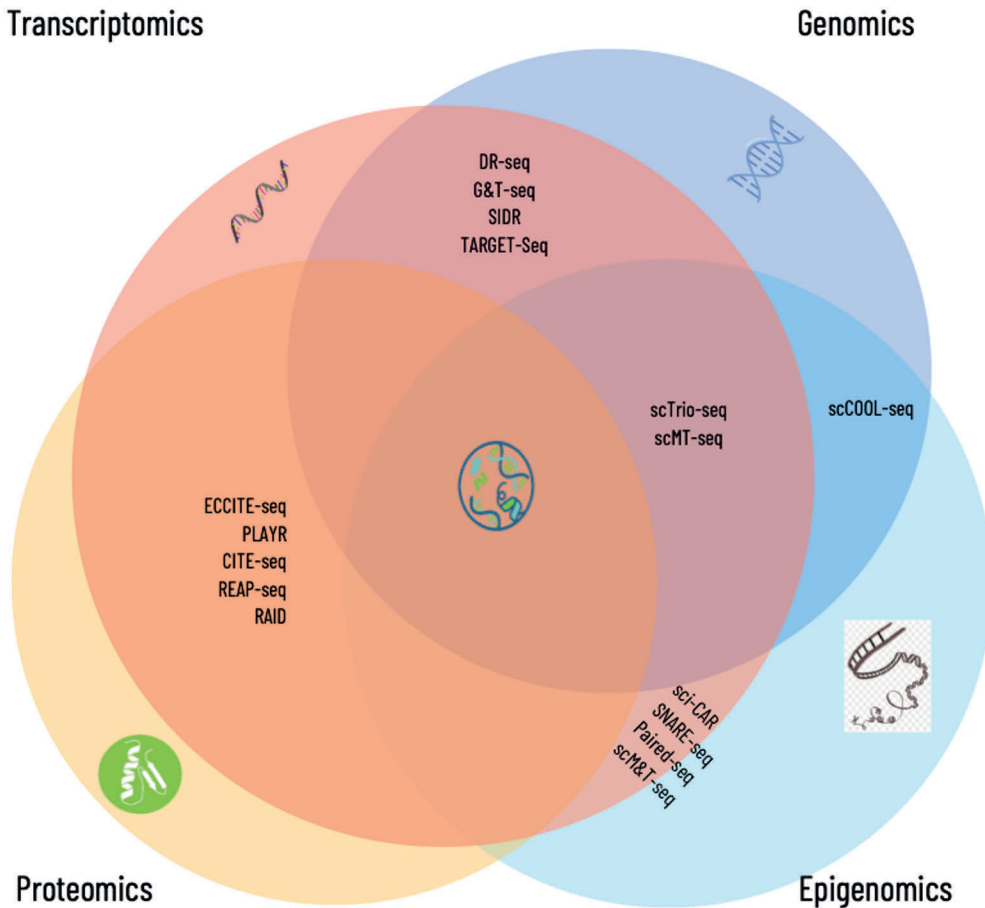


Figure 2. Single-cell multi-omics protocols. Single-cell multi-omics protocols such as DR-seq, G&T-seq, SIDR and TARGET-Seq profile transcriptomics and genomics data from a single cell. There are also other protocols i.e Sci-CAR, SNARE-seq, Paired-seq, and scM&T-seq that profile single-cell transcriptomics and epigenetics data. Additionally, protocols such as ECCITE-seq, PLAYR, CITE-seq, REAP-seq and RAID are capable of profiling the transcriptomics and proteomic data from a single-cell. ScCOOL-seq is also another protocol to profile genomic and epigenomic data from a single-cell. While most protocols profile two omics levels from a single-cell, there are few approaches where more than two omics data are profiled, for example, scTrio-seq and scMT-seq protocols are capable of profiling transcriptomics, genomics and epigenetic data from a single-cell.

On the other hand, the use of oligo-dT primer-coated magnetic beads for the separation of the polyadenylated mRNA from DNA has shown an efficient way of multi-modal profiling of mRNA and DNA molecules from a single cell (**Figure 2**) [72]. Once the transcriptomic and the DNA components are separated, both are amplified and sequenced separately. Unlike a physical separation strategy, another approach proposed by [73] has adopted a quasilinear amplification method to pre-

amplify the DNA and mRNA components from a single cell without physical separation. The pre-amplified components are then separated as the DNA and mRNA components for further amplification and sequencing.

2.1.2.2 Transcriptome and Epigenome

In multimodal profiling of single-cell epigenome and transcriptome, scM&T-seq [74] extended the utility of the separated genomic DNA and mRNA from G&T-seq [72] to profile single-cell methylation using the genomic DNA for bisulfite conversion. Other similar methods such as scMT-seq [75] and scTrio-seq [76] were also used to profile the DNA methylome, genomic DNA, and transcriptomes from a single cell (**Figure 2**). There have been also multiple protocols that profile different epigenetic layers from a single cell. scCOOL-seq [44] and scNOME-seq [77] were able to profile the joint epigenetic profiling for chromatin state and DNA methylation together (**Figure 2**). Adding a transcriptome profiling layer, scNMT-seq [78] and scCAT-seq [79] enhanced the former protocols.

2.1.2.3 Transcriptome and Proteome

The parallel profiling of proteome and transcriptome from a single cell has also got significant momentum recently. The recently developed method by [80] for simultaneous quantification of both protein and transcript utilizes the tagged-oligo labeling followed by the qPCR amplification. The Proximity Extension Assay (PEA) is the method that tags two antibodies that recognize the two epitopes of the same protein. This method is employed in simultaneous protein and targeted RNA profiling [80,81]. REAP-seq [82] and CITE-seq [83] protocols (**Figure 2**) make use of oligonucleotide-labeled antibodies to simultaneously readout surface protein and transcriptome measurements from a single cell [84]. Proximity Ligation Assay for RNA (PLAYR) [85], a mass cytometry-based method, attaches and ligates the RNA transcripts to the isotope-labeled probes so that the transcript abundance is measured simultaneously with elemental isotope-labeled protein. A single-cell RNA and Immunodetection (RAID) [86] is another reversible fixation-based protocol that uses Antibody RNA-Barcode Conjugates (ARCs) for simultaneous detection of intercellular phospho-protein together with transcriptomes from single cells.

2.1.2.4 Challenges and Opportunities of Single cell multi-omics

In general, technologies for incorporating multiple molecular profiling from a single cell overcome several challenges that remained unsolved in a single omics profiling from an individual level. For example, cellular heterogeneity analysis using only

scRNA-seq data has low sequence coverage resulting in inefficiently detecting the lowly expressed genes as a dropout. In addition to that, the downstream analysis, such as cell-type identification, is based on only a single molecular component (mRNA) in a cell, which gives a partial view or snapshot of the cell identity. However, the multi-omics single-cell technologies overcome these challenges by adding multiple layers of omics features or views to the cell-type identification analysis. Such an approach in return help researchers to understand detailed cellular function, cell-cell interactions and its implication in different biological processes and mechanisms.

Another important aspect of using a multi-omics approach at single-cell resolution opens the way for developing mechanistic models that can relate the interaction and the relationship among multiple layers of omics measurement (epigenetic variations, gene expression and protein expression) to unlock different molecular interplay within the cell. This enhances the study of gene expression dynamics and gene regulatory networks in a multi-factorial fashion. For example, [78] profiled the chromatin accessibility, DNA methylation, and transcriptome simultaneously from mouse embryonic stem cells and found novel links between the three molecular layers revealing the dynamics coupling the three omics layers in differentiating cells. It was also shown that CNVs cause proportional changes in RNA expression of genes within the gained or lost genomic regions from human hepatocellular carcinoma cells using single-cell triple omics sequencing [76]. The single-cell multi-omics technologies have also tremendous potential for clinical applications. In cancerous cells, tumor heterogeneity plays a crucial role in drug resistance, relapse, and metastasis [87]. Therefore, accurately identifying tumor subpopulations using a multi-omics approach enhances different biomedical and clinical applications including adaptive and precision medicine.

2.2 Single-cell RNA-seq Data Analysis

Single-cell RNA sequencing is one of the most used techniques to study the gene expression dynamics among the heterogeneous cell population. The single-cell RNA-seq data analysis has several upstream and downstream analysis steps (**Figure 3**). The upstream analysis includes quality control at the sequencing read level from fastq file, transcript quantification, and normalization, while the secondary analysis includes cell heterogeneity analysis, cell trajectory inference, marker gene identification, and gene regulatory network analysis. This chapter briefly summarizes the state-of-art workflow and tools for the analysis of single-cell RNA-seq data.

2.2.1 Quality control

The upstream analysis mainly focuses on ensuring the quality of the data from the sequencing machines and preparing the data for the secondary analysis. This includes the quality control on raw reads level and digital expression count matrix (**Figure 3**). The raw read level quality control on fastq files assures the quality of the reads from the sequencing machine. FASTQC which is a tool used for bulk RNA-seq data, is also widely used in read quality control for single-cell RNA-seq data. Other single-cell specific tools such as CellRanger [49], indrops [48], SEQC [88], or zUMIs [89] perform similar read level quality control in addition to demultiplexing, genome alignment, and gene expression quantification in an automated manner to produce the digital expression matrix.

This digital expression matrix also has to pass through different quality control steps. For example, for data generated by UMI-based library preparation, the cell barcode might mistakenly tag more than one cell (doublet) or it might not even tag any of the cells. On the other hand, there is a probability that a single cell might be tagged by multiple barcodes resulting in barcode multiplets [90]. Recent tools such as scrublet [91] and DoubletFinder [92], developed to address this issue, implement the simulation-based approach where doublets are simulated from the dataset itself and then the similarity between the real and simulated doublet is calculated to infer the doublets.

Another primary quality control step is to remove the cellular barcode that does not represent the actual cell. This is mainly done by defining the minimum threshold for the UMIs required to consider the given barcode as a cell and filtering out the barcodes that do not satisfy the criterion (**Figure 3**). The alternative to this method is to estimate the background amount of RNA in empty wells or droplets and then to keep the cell barcodes that significantly deviate from the background [93].

Even if the quantified RNA molecules are significant, there should be another layer of quality control on the detected genes/transcripts to assure that the cells are not damaged or dying cells. This could be achieved by considering the number of quantified genes/transcripts and the proportion of transcripts derived from the mitochondrial genome together with inspecting the proportion of unmapped and multi-mapped reads [94].

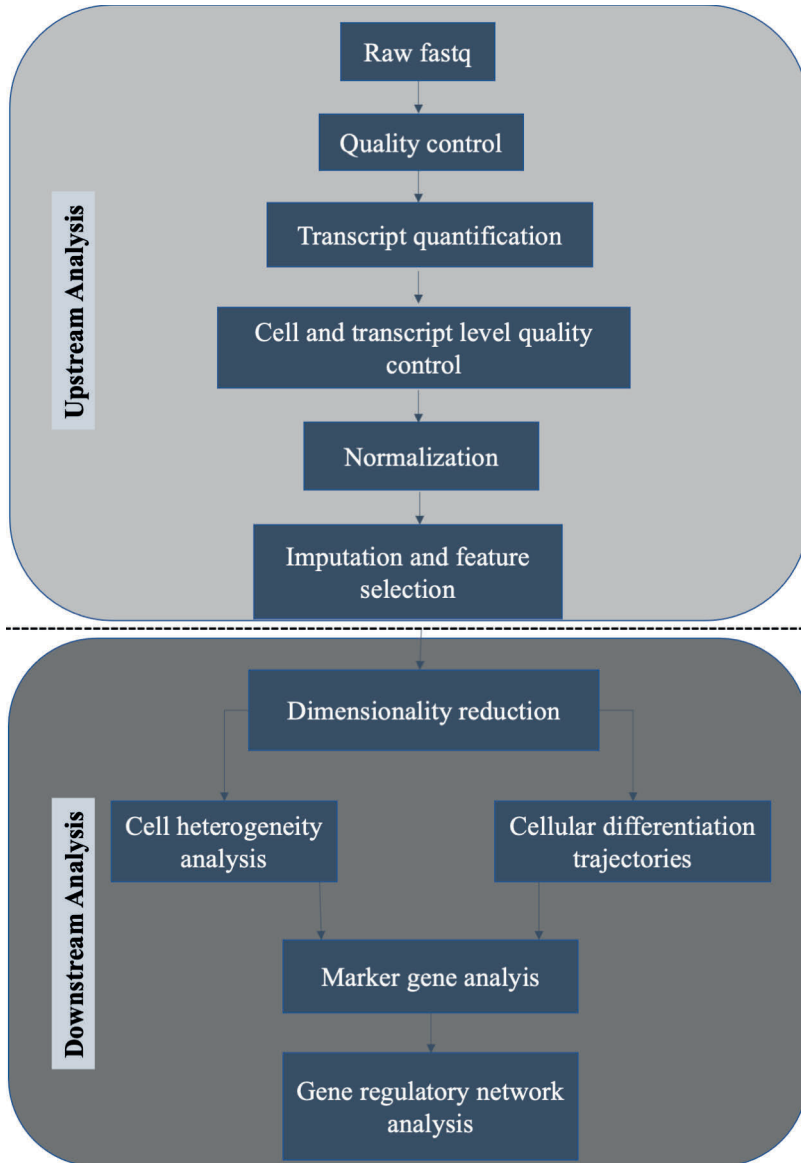


Figure 3. scRNA-seq data analysis workflow. The scRNA-seq data analysis has upstream and downstream analytical stages. The upstream analysis mainly focuses on quality control at raw fastq level and after read mapping for transcript quantification. Normalization, imputation and feature selection are also part of the upstream analysis. The downstream analysis comprises of the dimensionality reduction for cell heterogeneity analysis using clustering algorithms and cell differentiation analysis. Cell type specific marker gene analysis and cell type specific gene regulatory network analysis are also part of the downstream analysis.

2.2.2 Normalization

Even though the quality control at the raw read and UMI/read count under each cell add up to a certain level of data quality, different cells have different sequencing depths, where the number of reads in each of the cells varies. To account for this varying sequencing depth, the “size-factor” based normalization methods are used for bulk sequencing data such as RPKM and TPM, where the expression values are divided by the sequencing depth-specific size factor. This approach is also used for normalizing the single-cell RNA-seq data (**Figure 3**). However, a large number of zeros in the matrix are attributed to the amount of captured RNA molecules in a cell and with varying sequencing depth. Therefore, single-cell RNA-seq-specific normalization packages such as scran [95] use pools of cells for estimating size factor.

The lowly expressed genes in the zero-inflated expression in scRNA-seq data might behave differently than the highly expressed ones. The quantile regression-based normalization method, SCnorm [96], addresses this issue by accounting for the transcript expression on sequencing depth for each gene. Another method called transform [97] uses the cellular sequencing depth as a covariate in a generalized linear model to normalize the data. The Bayesian-based normalization method for scRNA-seq count data, bayNorm [98], accounts for the effect of the mRNA capture for scaling and normalizing the data [99].

Further, other confounding factors either technical or biological that add up unwanted variability to the dataset has to be removed. For example, the batch effects that arise from a different time of the experiment, the person experimenting, and differences in reagents and sequencing machine are technical confounders. Such batch effects have to be corrected computationally before the downstream analysis. One of the methods to achieve the batch correction in scRNA-seq data uses the mNN (mutual nearest neighbor) approach implemented in mnnCorrect [100]. The mnnCorrect [100] calculates the mutual nearest neighbor between cells in different batches to identify the common biological features across the batches. A similar approach was adopted for batch correction in Seurat2 [101] to find “anchors” for the canonical correlation analysis (CCA) projected cells. Several other tools such as Seurat3 [102], MMD-ResNet [103], Harmony [104], Scanorama [105], BBKNN [106], scGen [107], ComBat [108], LIGER [109], scMerge [110], and ZINB-WaVE [111] also recently implemented different methods to address the batch correction in scRNA-seq data.

In addition, biological confounding factors such as cycling cells bring another layer of unwanted variations among the cell population. Therefore, such confounding factors must be removed from the data analysis. Tools such as scPLS [112], RUV [113], and scLVM [114] use target and control genes to infer and remove the biological confounding factors. However, while it is important to remove

the cell cycle effect from the given dataset to increase the data quality for downstream analysis, computational categorization of cells into their cell cycling stages and distinguishing the cycling cells from quiescent cells also increase the efficiencies of sub-population detection and help to study the differences among the non-cycling and cycling subpopulations. Computational tools such as Seurat [101] utilize the known G1/S and G2/M cell-cycle marker genes to infer the cell-cycle stages. Another tool called Cyclone [115] uses a relative expression of pair of genes to infer the cells to G1, S or G2/M stages. Once the cells are assigned to their corresponding cycling stages, both tools use the linear regression model to regress out the differences [99].

2.2.3 Imputation and Feature Selection

The scRNA-seq expression data matrix is known for its sparsity, which is associated with low amounts of starting material, low RNA capturing and sequencing efficiencies of existing protocols, resulting in “dropout” events, where large proportions of genes in some of the cells get false zero counts (**Figure 3**). There have been several statistical methods, such as MAGIC [116], scImpute [116], SAVER [117], VIPER [118] DrImpute [119], SAVER-X [120], DCA [121], and DeepImpute [122], proposed for imputation of the dropout event. For example, MAGIC [116] uses the Markov transition matrix, a data diffusion-based method, to define kernel distance measures among cells. Another method, scImpute [116], implements a two-component mixture model to calculate the dropout probability and uses LASSO for the imputation of dropout values. SAVER [117] and VIPER [118] are imputation methods that are based on linear regression and non-negative sparse regression models. The consensus clustering-based method, DrImpute [116], first performs the consensus clustering of cells and then uses the average cell similarity values to impute the dropout events. Another imputation method, SCRABBLE [123], implements scRNA-seq data imputation by using bulk RNA-seq as a constraint. The deep neural network-based imputation methods, such as SAVER-X [120], DCA [121] and DeepImpute [122], have managed to learn the non-linear relationships and structures in scRNA-seq data. SAVER-X [120] integrates the deep autoencoder with the Bayesian models to impute scRNA-seq data, whereas DCA [121] is an imputation method based on deep autoencoder. Another similar method, DeepImpute [122], implements the divide-and-conquer approach using multiple sub-neural networks for imputation. The benchmarking studies [124,125] on several scRNA-seq imputation tools showed imputation improved the gene expression recovery that was observed in bulk RNA-seq data. However, it is also noted that imputation does not improve the results in downstream analysis, including in

clustering, trajectory inference and constructing gene regulatory networks. Therefore, it is recommended to use imputation in scRNA-seq analysis cautiously.

There are about 23,000 genes or dimensions for any mouse or human experimental dataset, out of which the current sequencing protocols, such as droplet-based and the more sensitive SMART-seq-based methods, can capture about 1,000 to 5,000 genes and 10,000 genes, respectively [126,127]. This shows the scRNA-seq data is sparse high dimensional data, where the distance calculation in the downstream analysis, for example in cluster analysis, becomes problematic due to the “curse of dimensionality”. In addition, biological signals can easily be hindered by technical noise. Therefore, extracting features that are only attributed to the biological signal is crucial. In this regard, one of the methods for feature extraction/selection is simply to take the genes that have a higher number of non-zero values [128]. Another strategy is to extract genes with higher variance across the cells [129]. Seurat [101] uses the variance-based non-parametric feature extraction approach using the mean and variance expression values. However, such feature selection methods have their limitation as they do not account for the genes in rare cell types because of their minimal contribution towards the total cell variability. To overcome this challenge, GiniClust [130], which uses the Gini index to quantify unequal distributions of the transcript showed to identify features in rare clusters/cell types.

2.2.4 Dimensionality Reduction

The dimensionality reduction further improves the impact of high dimensionality in the downstream analysis (**Figure 3**). For example, the linear dimensionality reduction methods called principal component analysis (PCA) and multidimensional scaling (MDS) are used to reduce the dimensionality of the feature-selected matrix by computing the linear projection of top eigenvectors from the covariance matrix of the high dimensional data. Then in PCA, the principal components (PCs) that better explain the variance of the high dimensional data are selected to perform the downstream analysis. The number of PCs used for downstream analysis is dataset dependent. As a result, methods such as the “elbow curve”, where a fraction of variance explained by each of the PCs are plotted for visual identification of the points where the curve makes a shape bending with no significant variance change for further increments in PCs, are seldomly used to infer the cutoff PCs. Furthermore, the selected PCs can be used in the nonlinear dimensionality reduction methods like *t*-distributed stochastic neighbor embedding (*t*-SNE) [131] and Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [132] to unlock the non-linear structure or topology of the given scRNA-seq data for visualization. However, the results from both *t*-SNE and UMAP are sensitive to the

hyperparameters given as input and they should be carefully selected. These two methods are widely used for the visualization of the scRNA-seq data. In general, the downstream analysis, mainly clustering, and trajectory inference, heavily depend on dimensionality reduction. Therefore, one must select the dimensionality reduction method of choice carefully.

2.2.5 Cluster Analysis

Different clustering algorithms are used for scRNA-seq cellular heterogeneity analysis. Tools such as Seurat [101] and SC3 [133] use PCA-based dimensionality reduction before proceeding to the downstream cluster analysis (**Figure 3**). The Seurat1 [133] implemented the k nearest neighbor (k NN) graph-based clustering method on the significant PCs extracted from the feature-selected scRNA-seq data matrix. Moreover, the next versions of Seurat2 and 3 [101] have adopted canonical correlation analysis (CCA) based projection method instead of PCA for joint clustering of multiple single-cell RNA-seq data across different samples or technology using mutual nearest neighbors (m NN) graph-based clustering. Seurat4 [134] implemented the weighted nearest neighbors (w NN) graph-based joint clustering method for integrative cluster analysis of multi-omics single-cell datasets. Another consensus clustering-based framework for single-cell RNA-seq data analysis, SC3 [133], implements the PCA for reducing the dimensionality of the cell-cell distance matrix on the feature selected data and applies k -means clustering on the first d PCs. It then constructs a consensus matrix from several k -mean clustering results to determine the final consensus clustering result using hierarchical clustering.

Another dimensionality reduction and clustering method used in the scRNA-seq analysis is nonNegative matrix factorization (NMF). It is a factor analysis-based method for extracting sparse and meaningful features and structures from high-dimensional data with low-rank approximation. Given $X^{g \times c}$, the scRNA-seq data matrix of g genes and c cells, NMF factorizes it as a product of two non-negative and low-rank base (H) and coefficient (W) matrices, $X \sim WH$. The coefficient matrix ($W^{g \times r}$) is with the dimension of g and the number of factor r and the base matrix ($H^{r \times c}$) is with the dimension of a number of factors r by the number of cells c . The factor r determines the number of clusters or the number of reduced dimensions. [135] has demonstrated the use of NMF in identifying the cell types in a wide variety of single-cell RNA-seq data. Tools such as CRNMF [136] implemented the NMF for scRNA-seq clustering by modeling the dropouts as a sparse matrix. ccfindR [137] also combined NMF with Bayesian modeling to analyze cell-type heterogeneity in the cancer microenvironment. cNMF [138] and SOUP [139] are additional tools that

have implemented the NMF as a method of choice for both clustering and dimensionality reduction.

The Bayesian probabilistic models such as Latent Dirichlet Allocation (LDA) [140] are widely used for topic modeling in the field of text mining for identifying the hidden topics within a corpus of documents. The documents are modeled by a Dirichlet distribution, where words with higher probability are observed more frequently in the document cluster of a similar topic. LDA has also been applied for simultaneous dimensionality reduction and cluster analysis in scRNA-seq count data as it shares discreteness and sparsity with document data in topic modeling. In this regard, tools such as DIMM-SC [141], CELDA [142], and CellTree [143] have adopted LDA for both dimensionality reduction and cluster analysis.

2.2.6 Trajectory Inference

scRNA-seq is widely used for studying cellular development and pseudotime lineage tracing in differentiating cells (**Figure 3**). Mostly, the cell trajectory inference algorithms use the lower dimensional embedding for visualization and formulating optimal trajectory lines in pseudotime. For example, tools such as Waterfall [144] use k -mean clustering on the lower dimensional PCA spaces to build a linear cell lineage trajectory. In the same way, TSCAN [145] also uses the PCA for dimensionality reduction to run the minimum spanning tree (MST) algorithm for cell lineage trajectory inference. Diffusion map [146] is a non-linear dimensionality reduction method that utilizes local similarity measures for creating a time-dependent diffusion process to re-order the high dimensional data according to the underlying geometry in the lower dimensional manifold. Such kinds of time-dependent diffusion process-based dimensionality reduction methods are suitable for analyzing the single-cell experimental data from differentiation experiments or time course scRNA-seq data. In this regard, destiny [147] and Wishbone [148] have implemented diffusion map to infer the cellular differentiation lineages from single-cell RNA-seq data in pseudo-time. Another non-linear dimensionality reduction method called locally linear embedding (LLE) [149] computes the k nearest neighbor for each of the data points to find the lower dimensional embedding by optimization of eigenvectors. LLE-based lower dimensional embedding has been utilized in SLICER [150] for projecting the scRNA-seq data into the lower dimensional space for reconstructing the cellular trajectory. Deep neural network-based autoencoders have recently got popularity in different domains including in scRNA-seq data analysis. scVI [151] has demonstrated the application of autoencoders for preprocessing and dimensionality reduction for scRNA-seq data. Tools such as Dhaka [152], scScope [153], VASC [154], and DCA [121] also implement deep neural network-based dimensionality reduction. In addition, Monocle [155] has

implemented UMAP and independent component analysis (ICA) as dimensionality reduction for constructing cell lineage trajectory [155].

2.2.7 Marker Gene Analysis

Identifying the heterogeneous cell type/clusters using any of the above-mentioned dimensionality reduction techniques followed by clustering and/or trajectory inference techniques, the next step is to find out molecular markers that give the group of cells/cluster the identity that it holds (**Figure 3**). Different differential expression analysis methods developed for bulk RNA-seq data such as ROTS [156], DESeq [157] and edgeR [158] are also used in the context of analyzing differentially expressed gene (DEG) in scRNA-seq data [159,160]. However, there have also been single-cell specific differential expression tools that account for single-cell RNA-seq properties such as dropout, higher technical and biological noises, and lower library sizes [121,149–160]. Most of these tools have their underlying model assumptions for scRNA-seq count data. For example, DESeq [157] and edgeR [158] assume the negative binomial model, while DEsingle [163] assumes the zero-inflated negative binomial model. Other tools such as BPSC [173], MAST [169], scDD [170], and Monocle [155,174] model the dropouts with mixture models. The nonparametric DEG implementations such as SigEMD [170], EMDomics [171], and D3E [172] utilize the distance metrics among the distribution of genes between two conditions for differential expression analysis.

A recent study showed that even though there are several single-cell specific DEG analysis tools, there has not been a gold standard and the DEG result depends on the underlying scRNA-seq data structure [165,166]. In addition, both bulk and single-cell-specific DEG tools are sensitive to batch effects and sample size [175]. Generally, accurate identification of marker or differentially expressed genes in scRNA-seq data remains challenging, and robust and accurate tools that account for the multimodality, sparsity, dropout and nature of scRNA-seq data are yet to be rolled out [165,166,175].

The same DEG methods that are used for the scRNA-seq clusters can be applied to identify markers during lineage differentiation in trajectory analysis. However, the DEGs that are obtained by comparing the cluster of cells obtained by the trajectory inference obscure the interpretation because of the pseudotime gene expression values are not at the same pseudotime point [176]. This creates complexity to adopt the tools that are already developed for bulk time-series RNA-seq data [177–179]. As a result, there have been single-cell RNA-seq-specific pseudotime lineage trajectory-based differential expression methods that consider the continuous expression resolution along the trajectory and compare expression differences across the lineages. In this regard, Monocle [155] and TSCAN [145]

have implemented an additive model to associate the gene expression and the differentiation of a linear lineage in a pseudo time. However, this method suffers from accounting for multiple lineages or bifurcating trajectories [176]. Another mixture model-based method that assumed each of the mixture components representing each lineage was implemented in Gpfates [180]. However, it cannot handle more than one bifurcation lineage. The later version of Monocle2 [103] implemented the branched expression analysis modeling (BEAM) method that used generalized linear modeling (GLM) [181] for analyzing the bifurcation or multifurcation gene expression difference along the trajectories or lineages. The software frameworks such as tradeSeq [176] also implemented the generalized additive models based differential analysis methods for multiple bifurcating trajectories.

2.2.8 Single-cell Gene Regulatory Network Analysis

Gene expression is regulated by the complex regulatory interactions among other genes and molecules in combination with chromatin accessibility, transcription factors, and other cellular microenvironments. Unlocking this complex gene regulatory network at single-cell resolution is useful to understand the interacting genes and the biological processes involved in different developmental or disease stages (**Figure 3**). This facilitates the discovery of disease biomarkers and identifies potential pathways and drug targets. The co-expression-based approach for constructing a gene regulatory network (GRN), which has been widely used in bulk RNA-seq data, can also be used for scRNA-seq data. The co-expression approach assumes that if two genes show co-expression, it is expected that they are in a regulatory relationship. For example, SCENIC [182], a single-cell specific GRN inference tool, identifies the potential TF target genes using co-expression analysis and then it performs the TF-motif enrichment analysis to measure the regulon activity at single-cell resolution. SINCERA [161] and NLNET [183] are also among other single-cell specific GRN inference tools that are based on co-expression analysis.

The other approach mostly implemented to construct the GRN inference at single-cell resolution implements the Boolean model. It uses a Boolean operator to indicate the relationship between nodes or genes (0 for unexpressed and 1 for expressed) and the edges show the gene's topology. The SCNS [184], BTR [185] and Boolean Pseudotime [186] are single-cell specific Boolean model-based GRN inference implementations. However, calculating the Boolean function is expensive in terms of computational cost and it poses a constraint on the scalability of such an approach [187]. Another approach uses differential equations to model the dynamics of gene expression as a function of the expression level of other genes or cellular

environmental factors. Inference Snapshot [188], SCODE [189], and SCOUP [190] utilize the pseudo-time inference algorithms for cell ordering and solve either ordinary differential equation (ODE) or stochastic differential equation (SDE) in order to construct GRN. As this approach is dependent on the pseudo-time inference, any error or noise introduced at this stage might affect the downstream analysis and potentially hinder the accurate network construction [187]. In general, GRN inference from scRNA-seq data is relatively new and the current methods and tools are sensitive to technical noises according to the comparative review studies [191].

2.3 Integrative Single-cell Multi-omics Data Analysis

The major aim of doing multi-omics cluster analysis is to understand the shared latent structure from multiple high-dimensional datasets to get a comprehensive understanding of the single-cell multi-omics dataset. One of the challenges in doing such multi-omics data alignment is that the dimension and target measurement of multi-omics data are different. Therefore, first the datasets have to be coordinated for the joint representation. Additionally, the high-dimensional nature of multi-omics dataset create challenge in constructing a common latent semantic representation across multiple datasets.

In order to address such a challenge, the Manifold alignment algorithm, which aligns disparate multi-omics dataset for discovering the underlying shared latent semantic structure is recently used for single-cell multimodal data integration. Manifold alignment algorithms find the lower-dimensional embedding for multiple datasets simultaneously by inferring correspondence information among each of the manifolds in the multiple lower-dimensional embedding. There are two major steps in manifold aligning algorithms: one is finding the intrinsic relationships of features within each of the datasets by extracting the underlying low-dimensional representation of the local geometry as a manifold using a graph Laplacian associated among each of the datasets. The second step is mapping this lower-dimensional embedding into the joint latent space so that locally similar instances within each dataset and the corresponding instances across datasets are close or identical in joint space.

Canonical Correlation Analysis (CCA) is another multivariate analysis method to examine the relationship between two sets of datasets based on their correlation. It determines the set of linear combinations of all variables in each of the two datasets in such a way that maximizes the correlation between the two linear datasets best explaining both within and between dataset variability. This basic CCA can be enhanced in order to accommodate different datasets. For example, the high dimensionality nature of the multiple omics data and insufficient sample size may

pose constraint for a linear combination of all the features leaving poor biological interpretability of the CCA results. In order to combat this challenge, the CCA variant called sparse CCA [192] which finds the sparse loadings that maximize correlation between the subsets of variables using lasso penalty based on SVD (Singular value decomposition), is proposed by [193].

Non-negative Matrix Factorization (NMF) is a factor analysis-based method for extracting sparse and meaningful features of the high dimensional data using low-rank approximation. There are few approaches that suggested the adoption of NFM for the multi-omics data clustering. One of the approaches is using a multi-view version of Frobenius norm optimization for finding optimal common coefficient matrix.

AutoEncoder is one of the unsupervised generative deep neural networks with an architecture of input, hidden and output layers with the bottleneck in the middle of the hidden layers indicating the most compressed transformation of input data. The hidden layer consists of two parts: encoders and decoder layers. The encoder compresses the input data so that it stores the compressed or lower dimensional representation of data at the bottleneck layer whereas the decoder part decompresses the data at the bottleneck layer to regenerate the original high dimensional input data. The compressed data at the bottleneck layer removes the noise in the original input layer and represents the lower dimensional representation of the input data, hence the compressed data can be, for example, used for further cluster analysis using any conventional clustering algorithms.

2.4 Dirichlet Process Mixture Models for Cluster Analysis

Cluster analysis is a method by which similar high-dimensional data are grouped together and dissimilar ones are grouped separately as independent clusters. Bayesian Dirichlet admixture models are mainly used in cluster analysis for topic modeling. The parametric Dirichlet admixture model called latent Dirichlet allocation (LDA) utilizes the Dirichlet distribution as a priori, while its nonparametric counterpart, the Hierarchical Dirichlet process (HDP), uses the Dirichlet process as a model prior. Both models are widely used in natural language processing (NLP) for cluster analysis [194]. The identification of cellular heterogeneity from single-cell omics data is mainly based upon the cluster analysis. Hence, **Article I** explored the utility of the Bayesian clustering methods such as LDA and HDP in the context of cell heterogeneity analysis in scRNA-seq data.

2.4.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet allocation (LDA) [194] is a finite parametric admixture model for the clustering task. Mostly, LDA is used for topic modeling where it assumes a corpus with a collection of documents with finite topics. The topic distribution is assumed to have a Dirichlet prior over the finite number of topics. In addition, each topic is characterized by the distribution of the words [195]. A graphical representation of LDA is presented in **Figure 4**.

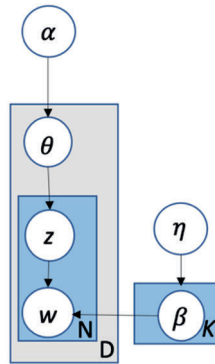


Figure 4. Graphical representation of LDA model for topic modeling. α is the priori concentration parameter of document-topic distribution while η is the priori concentration parameter for word-topic distribution. β is the word-topic distribution for K topics and θ is the document-topic distributions for D documents. z is the word-topic assignment for observed word w .

The generative graphical representation of the LDA model in **Figure 4** represents D documents with N words for a given document d . K is the total number of topics and α is the concentration parameter for the symmetric Dirichlet distribution which is used as a priori for the document-topic distribution. η is also a parameter for the symmetric Dirichlet distribution for the topic-word prior distribution. θ is a topic distribution over D documents while β is a word distribution of K topics. Z is the topic assignment for each word in a document [194,195]. The two mostly used inference algorithms for computing the posterior distributions are the Gibbs sampler and variational inference [141–143,196,197].

2.4.2 Hierarchical Dirichlet process (HDP)

HDP is a nonparametric generalization of LDA with countably infinite numbers of components or clusters given as prior. It uses the Dirichlet process (DP) as a priori for model construction. As a result, unlike the LDA, the number of clusters are not

a predefined model parameter, rather it is inferred from the given dataset using a posterior inference algorithm.

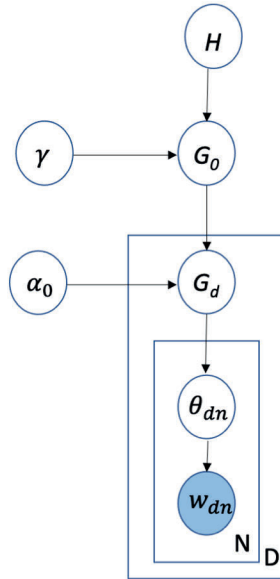


Figure 5. Graphical representation of the HDP model for topic modeling. The Dirichlet process G_0 with H base distribution and γ priori concentration parameter is nested in another DP G_d as a base distribution and together with α_0 as priori concentration parameter. θ_{dn} is the topic distribution for document d with n words while w_{dn} is the observed word for document d with n words where D is the total number of document and N is the total number of words in each document.

The HDP graphical model representation in **Figure 5** illustrates the model prior G_0 is drawn from a Dirichlet process (DP) with concentration parameter γ and base distribution H to construct document distribution G_d according to the Dirichlet process for clustering D documents having N words. At the same time, G_d is used as the base distribution with the initial concentration parameter α_0 to construct the word distribution θ_{dn} according to the Dirichlet process. As the draw from a DP is demonstrated to be discrete according to the stick-breaking construction [198] both θ_{dn} and G_d are discrete, leading to corpus level word topic clustering and document level topic clustering.

2.4.3 Inference Methods

Inference algorithms are used in both LDA and HDP to get full posterior distribution. The two mostly used inference methods are Markov chain Monte Carlo (*MCMC*) [199] based algorithms, such as Gibbs sampler, and variational inference. *MCMC* is

one of the techniques used for nonparametric inference [199] for estimating the posterior distribution of N samples from the given distribution by ergodic averaging. Gibbs sampling, *MCMC*-based technique, generates the posterior samples by swapping through each variable to sample from its conditional distribution fixing the remaining variable to their current constant till convergence [199]. The initial sample assignment is random, as a result, samples simulated at the initial stage of the iteration are not representatives of the actual posterior distribution. *MCMC* algorithms are expected to run for quite a large number of iterations to converge to the target posterior distribution. The non-representative samplings from the early stage iterations are discarded [199]. *MCMC*-based sampling technique is limited to small-scale samples as it is computationally expensive. An alternative inference method that can scale to a larger dataset is variational inference [200]. Variational inference [201] is an inference method that approximates the posterior distribution by optimization. It works in such a way that it first posits a family of densities and then finds a member of that family close to the target. Closeness is measured by Kullback-Leibler divergence [201]. Compared to *MCMC-based* sampling methods, variational inference tends to be faster and easier to scale to large data.

2.4.4 Application of DPMM for Single-cell Omics Clustering

The intuition of topic modeling, where the document's topic distribution in a corpus and word's topic distribution is used to cluster documents by their topics, is analogous to the concept of cell-type clustering from single-cell omics data and finding out the markers for each of the cell-type specific clusters. In the context of single-cell RNA-seq, the entire single-cell dataset can be considered as a corpus with read counts/UMI counts in each of the cells as word counts in the context of topic modeling. Therefore, LDA/HDP models can be used to cluster cells of similar types and genes or molecular markers of specific omics data at the same time. In this regard, studies have shown the use of LDA in the context of clustering scRNA-seq expression data [141–143,196]. Another study [202] also implemented LDA for the simultaneous discovery of cell types together with the enhancer and relevant transcription factors from differentiating hematopoietic single-cell ATAC-seq data. HDP is applied for regulatory network segmentation and clustering bulk gene expression data [203]. HDP also has shown a potential to improve cell-type clustering by correcting technical variation with cell-specific scaling in scRNA-seq data [197].

HDP model is the non-parametric counterpart of the LDA model whose prior model parameters are drawn from the Dirichlet process with countably infinite model components. As a result, HDP does not require a pre-defined cluster number as an input parameter, whereas the LDA model accepts the predefined number of clusters

as an input parameter. The application of both LDA and HDP models for scRNA-seq data has shown improved cell-type clustering results [141–143,196,197]. However, the comprehensive comparison between the LDA and HDP models for clustering scRNA-seq data has not been assessed thoroughly.

2.5 Single-cell Transcriptomics of Endometrium

Endometrium is the outer layer of the uterus where the embryo implants in the initial stage of pregnancy. There are different cell-types that facilitate proper differentiation (decidualization) of endometrium for successful placental formation and pregnancy. Single-cell transcriptomics studies have enabled transcriptomic atlas of heterogeneous endometrial cells from both fetal and maternal perspectives. In this respect, [204] studied the cellular heterogeneity from samples taken from placenta, decidua and blood of the 1st trimester pregnancy selectively terminated between week 6 and week 14 of gestation. Another study [205] profiled the transcriptional landscape of placental villous trees, chorionic membranes and basal plate from woman at term pregnancy. Additionally, [206] also demonstrated the single-cell transcriptomic dynamics of endometrium across the menstrual cycle, providing novel insights towards endometrial transformation during the menstrual cycle. Such studies and the publicly available datasets opened an opportunity to further investigate the impacts of different endometrial cell-types on pregnancy disorders such as preeclampsia. Further studies on endometrial cell-type specific gene regulatory network analysis also have the potential to reveal the unknown cellular processes in pregnancy disorder.

3 Aims

This thesis work focuses on answering the following methodological and biological research questions:

- I. How applicable and efficient are the Dirichlet process mixture models such as LDA and HDP for clustering single-cell RNA-seq data and how they perform on different single-cell cell RNA-seq data using intrinsic and extrinsic cluster quality measures?
- II. What insight can the integration of single-cell and bulk RNA-seq data give on the cell-type specific marker gene contributions for late- and early onset preeclampsia?
- III. What cell-type specific gene regulatory networks and TFs regulate decidual stromal and natural killer cell subpopulations during 1st trimester pregnancy?
- IV. What methodological approaches can efficiently be used for the integration of single-cell multi-omics data?

In the first research question, the study addressed the applicability and efficiencies of the two different Dirichlet process mixture models latent Dirichlet allocation (LDA) and hierarchical Dirichlet process (HDP) models for cluster analysis in scRNA-seq data. The study also compared the existing LDA-based tools designed for scRNA-seq data (**Article I**). The second research question answers the cell-type specific markers' contributions to the pregnancy complication in preeclampsia by integrating the existing bulk RNA-seq data from the early and late-onset preeclampsia with the recently generated scRNA-seq data (**Article II**). The third research question investigates the gene regulatory networks at single-cell resolution for the decidual stromal and natural killer cell subpopulations (**Article III**). The fourth research question is about inferring an efficient analytical approach and identifying the statistical methodologies for integrating and analyzing the single-cell multi-omics dataset (**Article IV**).

4 Materials and Methods

4.1 Datasets

4.1.1 Human Artificially Mixed Immune Dataset

1184 human single-cell transcriptomics datasets were artificially mixed for validating models used in **Article I**. These datasets consist of conventional dendritic cells, fibroblasts, lymphoblasts, B cells, CD4⁺ cells, and CD8⁺ cells that were collected from public repositories.

4.1.2 Mouse Cell Atlas Dataset

The single-cell transcriptomics samples taken from kidney and pancreas were extracted from three months aged mice. The scRNAseq data was collected from GEO with accession of *GSE109774*. These datasets were used to validate the clustering models used in **Article I**.

4.1.3 Human 1st Trimester Pregnancy Data

The publicly available single-cell transcriptomics data on 1st trimester pregnancy was downloaded from ArrayExpress with accession *E-MTAB-6701*, including six decidua samples (6 to 12 weeks' gestation). A total of 36,186 cells from healthy donor [204] were used in both **Article I** and **II**. Specifically in **Article II**, the single cell data from 1st trimester pregnancy, menstrual cycle and term pregnancy was analysed together with previous bulk preeclampsia data. In **Article III**, 12,584 decidual stromal cells with decidual stromal cell 1 (dS1), decidual stromal cell 2 (dS2) and decidual stromal cell 3 (dS3) annotations and 11,881 decidual natural killer cells with decidual natural killer proliferative (dNK p), decidual natural killer 1 (dNK1), decidual natural killer 2 (dNK2) and decidual natural killer 3 (dNK3) annotations were selected for the study.

4.1.4 Human Menstrual Cycle Data

The single-cell transcriptomics dataset from menstrual cycle study (cycle days 16-26) with a total of 71,032 cells from ten endometrial samples were extracted from public repository with GEO accession *GSE111976* [206] and used in the study published in **Article II**.

4.1.5 Human Term Pregnancy Data

The single-cell transcriptomics term pregnancy data were extracted with consent from dbGaP with accession phs001886.v1.p1. The study in **Article II** used 13,730 cells selected from no labor samples from basal plate and chorioamniotic membranes [205].

4.1.6 Pregnancy disorder datasets

The bulk transcriptomic profile for with severe early ($n=3$) and late ($n=3$) onset preeclampsia with normal ($n=3$) control samples was downloaded from [207] and used in **Article II** and **Article III**. In **Article II**, the differentially expressed genes between the late onset and normal samples and early onset and normal samples were used to study cell-type specific marker gene contributions from scRNA-seq data. In **Article III**, these differentially expressed genes were used to study the association of decidual stromal (dS) and natural killer (dNK) cells subpopulation specific regulon targets with the LOP and EOP. Additionally, the differential expression genes for the samples of the late secretory menstrual cycle (days 22–32) endometrium from women with previous severe preeclampsia ($n=17$) and the controls ($n=12$) was collected from [208] and used in **Article II** to study cell-type specific marker genes contribution from scRNA-seq data. The dNK up and down regulated genes in recurrent pregnancy loss (RPL) transcriptomics data from [209] and the unexplained RPL specific gene lists from [210] were used in **Article III** to study the contributions of dNK subpopulation specific regulons in pregnancy disorder.

4.1.7 Human cisTarget Motifs

The human cisTarget gene-motif ranking databases, 10 kbp up and downstream of transcription start site (TSS) together with 500 bp upstream and 100bp downstream of the TSS, was downloaded from iRegulon (gene-based motif rankings) (<https://resources.aertslab.org/cistarget/>) and it was used as a motif search space for TF-motif enrichment analysis in **Article III**.

4.2 Methods and Analytical Workflow

4.2.1 Comparison of LDA and HDP Models for Single-cell RNA-seq Clustering (Article I)

The analysis of single-cell RNA-seq data for cell-type identification mainly uses dimensionality reduction together with unsupervised clustering methods. In this respect, conventional clustering methods, i.e., distance-based, density-based, and graph-based clustering methods are commonly used in the bioinformatics community [211,212]. **Article I** focused on the Dirichlet process-based Bayesian mixture models, namely Latent Dirichlet Allocation (LDA) [194] and Hierarchical Dirichlet Process (HDP) [198] for clustering cells based on their gene expression. These methods have proven their performance in the field of natural language processing (NLP) for topic modeling [213]. Recent studies [196,202] also showed adopting LDA for clustering for single-cell omics data. Though the parametric LDA model has been suggested for single-cell transcriptomic and epigenetics cell clustering, there has not been comprehensive comparison of these models in the context of clustering cells from scRNA-seq data (**Figure 6**).

Moreover, inappropriate choice of the number of clusters as input parameters for clustering algorithms may impede the discovery of novel cell states or types. To address these challenges, the utility of the hierarchical Dirichlet process (HDP) for clustering scRNA-seq data as a non-parametric counterpart of LDA was investigated. Additionally, the performance of both methods using intrinsic and extrinsic cluster quality metrics was compared. The intrinsic cluster quality measures, like Davies-Bouldin index (*DB-index*) [214] and Calinski–Harabasz index (CH-index) [215], evaluate the intra-cluster compactness and inter-cluster separation as a criterion for cluster evaluation [216], whereas the extrinsic cluster quality measures such as Adjusted Rand Index (ARI) [217] evaluate the given clustering result in comparison with a reference clustering [218].

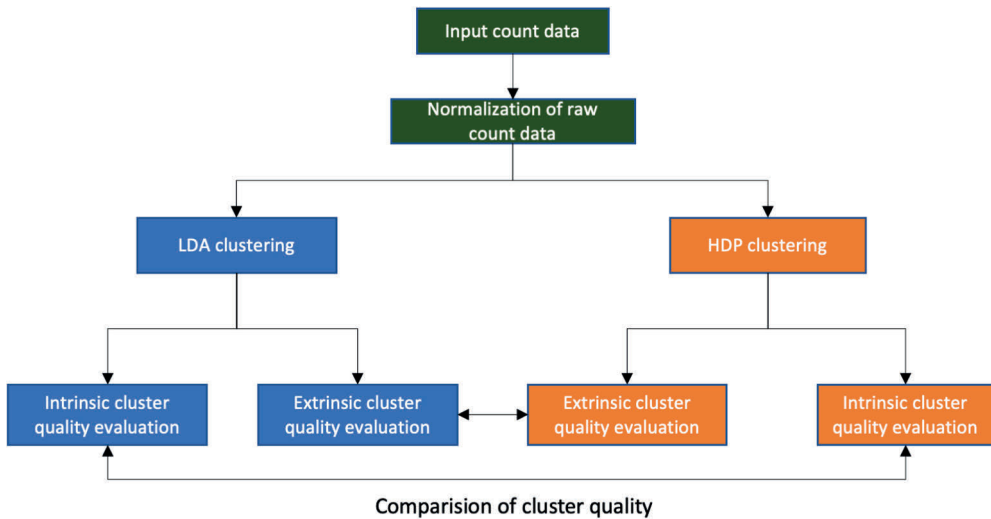


Figure 6. The workflow for comparison of LDA and HDP clustering models for single-cell RNA-seq data.

The python implementation for LDA and HDP models from the Gensim package was used for clustering each of the scRNA-seq data. The workflow for comparing LDA and HDP started with library size normalization of the raw count data and rounding it to the nearest integer values, as both models are meant to integer count data. The input parameters for the LDA model were the number of clusters and model prior concentration parameters (α and η , **Figure 4**), while the input parameters for the non-parametric HDP model were the model priori concentration parameters (α , γ and H , **Figure 5**). For the sake of simplicity and to avoid multi-parameter optimization, default fixed concentration parameter ($\alpha=1$, $\gamma=1$ and $H=0.01$) was used for both models to compare the performance of the models in terms of cluster quality. The HDP clustering was repeated 20 times for each dataset. Similarly, 20 repetitions of LDA clustering for each dataset with an increasing number of clusters k from 2 to 20. The online variational inferences method was used for posterior inference in all experiments. Finally, the clustering results were compared using the *DB-index*, an intrinsic cluster quality metric and ARI, the extrinsic cluster quality measure using the reference cluster.

4.2.1.1 Measures of Cluster Quality

The clustering quality is assessed using intrinsic and extrinsic cluster quality measures. The intrinsic cluster quality measures involve compactness and separation as a criterion for cluster evaluation [216], whereas the extrinsic cluster quality

measures evaluate the overall clustering in comparison with reference clustering [218].

Davies-Bouldin index (DB-index) [214] is an intrinsic cluster quality metric, which uses the intra-cluster variance and inter-cluster separation to evaluate cluster quality. For a clustering result that partition data points into k clusters with each cluster having a centroid c , the *DB-index* is given by:

$$DB = \frac{1}{k} \sum_{i=1, i \neq j}^k \frac{\max(D_i + D_j)}{d(c_i, c_j)} \quad (1)$$

here k is the total number of clusters, D_i and D_j are the average Euclidian distance between all the data points in cluster i and j respectively to their cluster center and $d(c_i, c_j)$ is the distance between the i^{th} and j^{th} cluster centers c_i and c_j . The *DB-index* is then the summation of the maximum average distances between any of the two clusters, normalized by the distance between their corresponding cluster centroids. The smaller the *DB-index*, the more compact the data points in each of the clusters with their cluster centers apart from each other. Therefore, clustering results with the lowest *DB-index* have higher cluster quality.

Calinski-Harabasz index (CH-index) [215] is another intrinsic cluster quality measure that applies a minimum within-cluster sum of a square as a criterion. It is a variance ratio-based index defined by the ratio of the overall between-cluster variance to overall within-cluster variance.

$$CH - index = \left[\frac{\sum_{k=1}^K n_k \|c_k - c\|^2}{K - 1} \right] / \left[\frac{\sum_{k=1}^K \sum_{i=1}^{n_k} \|d_i - c_k\|^2}{N - K} \right] \quad (2)$$

Where K is the total number of total clusters, N is the total number of data points, d_i is the i^{th} datapoint, n_k is the number of data points in the cluster k , c_k is the cluster centroid of k^{th} cluster and c is the global centroid. A higher *CH-index* indicates that the clustering results are optimal.

Adjusted Rand Index (ARI) [217] is an extension of *Rand index (RI)* [219], which is used as an extrinsic measure of cluster accuracy by calculating the percentages of correct clustering for a given clustering assignment with respect to the reference cluster where I and J are the total number of clusters in a clustering result to be evaluated and the reference cluster respectively. ARI uses hypergeometric distribution as the model of randomness over the matching or contingency table M . Then the Adjusted rand index is given by:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{[\sum_i \binom{a_i}{2}] [\sum_j \binom{b_j}{2}]}{\binom{n}{2}}}{\frac{1}{2} [\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - \frac{[\sum_i \binom{a_i}{2}] [\sum_j \binom{b_j}{2}]}{\binom{n}{2}}} \quad (3)$$

Where, a_i is the number of data points in clustering result to be evaluated with i^{th} cluster, b_j is the number of data points in the reference cluster with j^{th} cluster and n_{ij} is the number of shared data points in both clustering result and reference cluster. n is the total number of data points. The ARI values range from 0 to 1, where values closer to 1 indicate higher clustering quality.

4.2.2 scRNA-seq Cluster Analysis of Decidual Cells (Article II & Article III)

Three single-cell RNA-seq datasets used in **Article II** (1st trimester pregnancy [204], menstrual cycle [206] and term pregnancy [205]) were collected from public repositories. Standard scRNA-seq data preprocessing and analysis workflow was used. The raw SRA files were extracted for the term pregnancy data and the primary analysis was executed using cellranger -3.1.0 with reference genome (hg38) and Seurat4. The UMI count data was extracted for the 1st trimester and menstrual cycle datasets. After the standard quality control and normalization using Seurat4, the cluster analysis was performed for each of the datasets using the “*FindClusters*” Seurat function with default parameters. Cell-type marker genes were extracted for all the three datasets using “*FindMarkers*” function in Seurat4 and combined to perform over-representation analysis with the previous bulk study results. The clusters were assigned based on the predefined canonical markers from the literature for each of the corresponding cell-types. Trajectory inference was done using the “*infer_trajectory*” function of Slingshot in the “dyno” version '0.1.2 [220]. The “*plot_dimred*” function was used for visualizing the subcellular differentiation trajectories and gene expression.

In **Article III**, the standard workflow for single-cell RNA-seq analysis using Scanpy [v1.8.2] was used for the reanalysis of the subcellular heterogeneity study. The UMI count-based quality control was done by filtering out cells with less than 200 detected genes and removing genes that were expressed only in less than three cells from the data matrix. The data were log-transformed after the library size normalization, and the highly variable genes were selected using 0.25 and 3 as a minimum and maximum mean respectively with the dispersion parameter of 0.5. The batch effects arising from the individual donors were accounted by using “*mnn_correct*” Scanpy function over the highly variable genes. The downstream

analysis started with selecting the optimal number of principal components (PCs) for neighborhood graph construction using the “*Elbow*” method. The first 50 PCs were found to be optimal to construct the neighborhood graphs with 20 neighbors. Further, the community detection algorithm “*leiden*” with resolution parameter of 0.3 was used to cluster the cells. The clustering results were reannotated according to the cell-type marker genes in such a way that elucidated the transcriptional and gene regulatory activity scores.

4.2.3 Gene Over-representation Analysis (Article II and Article III)

In **Article II**, the Fisher’s exact test using R was applied to study the cell-type specific marker genes contributions from scRNA-seq data and previously identified up and down regulated genes from bulk RNA-seq studies [207] for late and early onset preeclampsia. Additionally, the web-based software tool METASCAPE with Fisher’s exact test was used for gene-list over-representation analysis in **Article III** to study the association of the subpopulation specific TF target genes and their gene ontology terms.

4.2.4 Single-cell Gene Regulatory Network Analysis (Article III)

The gene regulatory network analysis in **Article III** was performed using the python implementation of pySCENIC [version 0.11.2] in three phases (**Figure 7**). In the first phase, the raw count matrix was used as an input to calculate the adjacency matrix using “*grnboost2*” function to predict the TF and their targets based on the co-expression analysis. Then in the second phase, “*ctx*” function was used for TF-motif enrichment analysis for regulon predictions. In the third phase, the “*aucell*” function with “*auc_threshold*” value of 0.01 was used to get the regulon specificity score matrix. The “*aucell*” function scores the enrichment of regulons as an area under the recovery curve (AUC) across all gene rankings in a cell. Then, the binarize function was used to specify regulon activity as a binary outcome (1 for on and 0 for off) in each of the cells. Finally, we used the R package “*ComplexHeatmap*” version 2.6.2 with “*heatmap*” function for plotting the heatmap of binary regulon activity score for the first top 10 regulons ranked by the regulon specificity score among each subpopulation. “Cytoscape” version 3.9.1 was used for the TF-target network visualization of specific regulons of interest. The simplified network visualization filtered the regulon target gene lists based on the upregulated subpopulation markers with a significance value of $FDR < 0.01$.

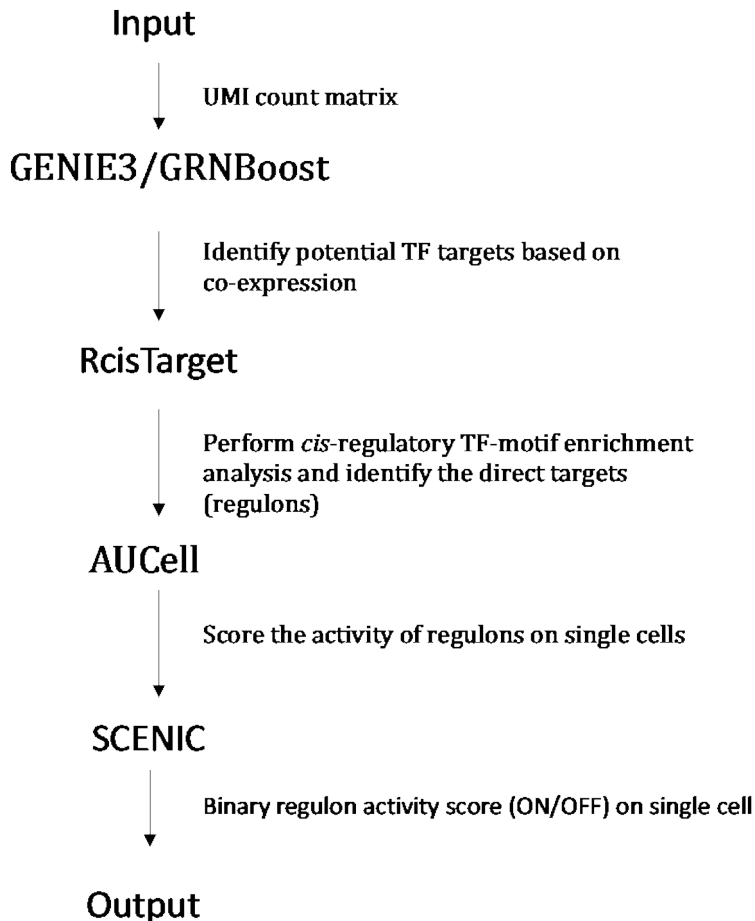


Figure 7. Single-cell gene regulatory network analysis workflow using SCENIC. Initially, The R packages GENIE3 [221] or GRNBoost [182] were used to identify TF and their potential target in a co-expression module. Then, RcisTarget is used for TF-motif enrichment analysis for identifying the direct targets or regulons. Finally, AUCell is used to score the activity of each regulon in a single-cell and this activity score is further binarized signifying the ON/OFF activities of the given regulon in a cell.

4.2.5 The Review of Single-cell Multi-omics Data Integration (Article IV)

The single-cell multi-omics technologies are relatively new techniques that were getting more attention at the time of the manuscript preparation. **Article IV** reviews the different strategies and methodological aspect of analyzing the single-cell multi-omics data from the literature. The state-of-art computational tools for the integrative analysis of single-cell multi-omics data were also summarized together with the challenges and opportunities in relation to the single-cell multi-omics data management and analysis.

5 Results

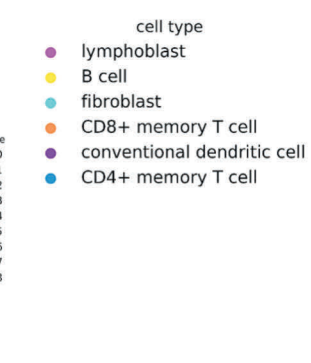
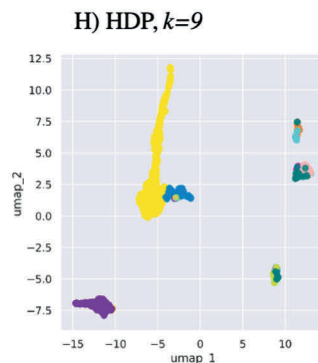
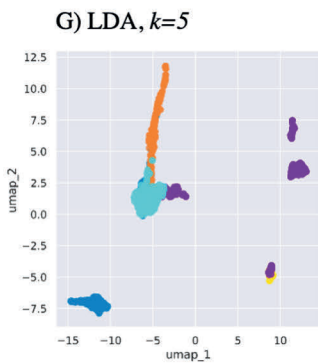
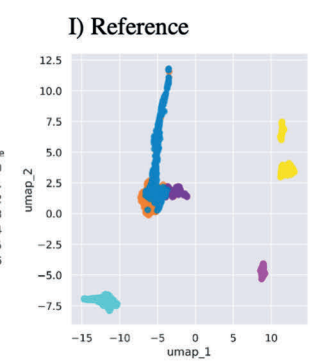
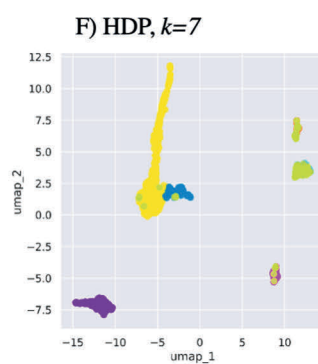
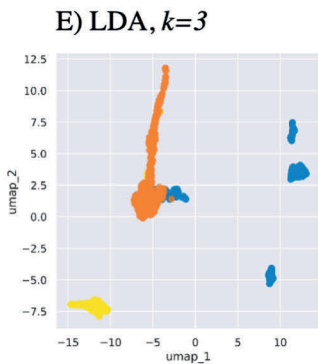
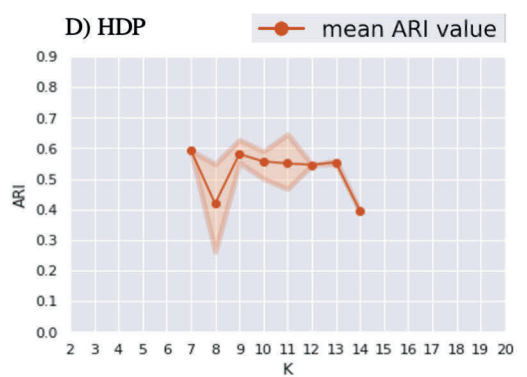
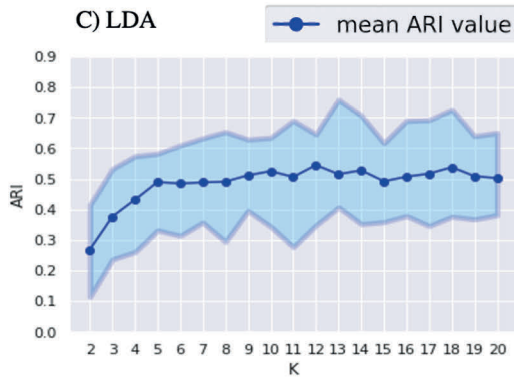
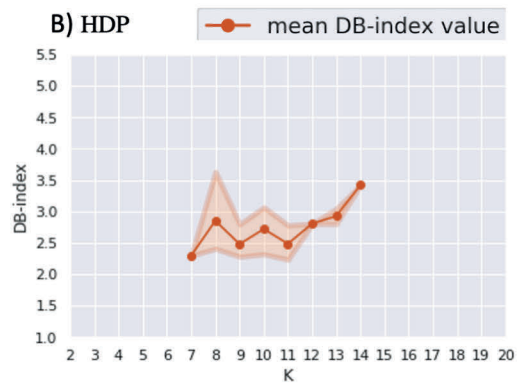
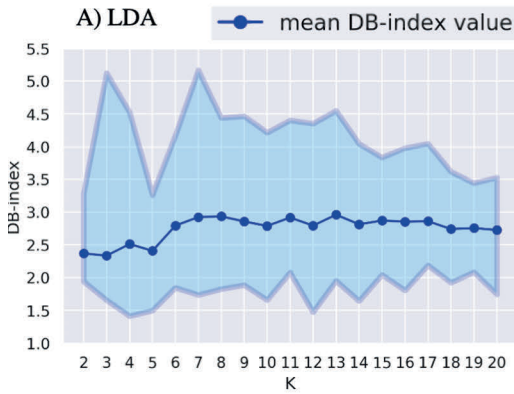
5.1 Comparison of LDA and HDP for scRNA-seq Clustering

LDA and HDP models for scRNA-seq cluster analysis have been adopted in recent years [141–143,196]. However, a comprehensive comparison between the LDA and HDP models for clustering scRNA-seq data has not been done. Generally, one of the advantages of HDP over LDA is that it does not require the number of clusters as an input parameter. In **Article I**, a comprehensive comparison of LDA and HDP models for clustering scRNA-seq data using both intrinsic and extrinsic cluster quality measures was conducted.

5.1.1 Clustering Performance

The LDA and HDP clustering performance was assessed using *DB-index*, ARI, and visual inspection at lower dimensions in four datasets. Overall, the clustering performances of LDA and HDP were dataset dependent. In the human immune cell dataset, HDP was able to discover known cell types in a slightly better way than LDA. LDA showed the minimum average *DB-index* values with cluster numbers $k=3$ and $k=5$, while the HDP clusters had the lowest average *DB-index* for $k=7$ and $k=9$ (**Figure 8**). Comparing this result in terms of ARI, HDP clustering resulted in a higher average ARI value (0.6) for $k=7$, whereas the highest average ARI value for LDA was 0.5 for $k=5$.

Figure 8. ► LDA and HDP comparison for clustering artificially mixed human immune cells. (A) and (B) uses intrinsic cluster quality measure *DB-index* in the y-axis and the number of clusters in the x-axis, while (C) and (D) uses the extrinsic cluster quality measure called ARI. The lines in the middle of figure shows the average *DB-index* and ARI values for the 20 times repeated experiments for both LDA and HDP clustering in the y-axis for the number of clusters in the x-axis. Similarly, the upper and the lower lines indicate the highest and the lowest *DB-index* and ARI values in the y-axis for the cluster number in the x-axis. (E-H) illustrates the 2D lower dimensional visualization of clustering results for selected high-quality clustering results from both LDA and HDP. (I) shows the visualization of the reference clustering. Figure is reproduced from Article I.



The other three datasets suggested that HDP itself does not perform better than LDA. However, the clustering results were relatively comparable. Generally, the results suggested that the HDP-based k value may be useful to guide the selection of the k value for LDA.

5.1.2 Computational Scalability

The experiment was performed in small to medium-sized single-cell RNA-seq data on 48-core Ubuntu 16.04 EC2 cloud instance. The run-time for a single analysis for LDA clustering in the immune cells (~1000 cells), pancreas cells (~2000 cells) and kidney cells (~3000 cells) took ~2-3 minutes, while the HDP clustering took ~6-28 minutes (**Table 1**). LDA and HDP run times for the decidua/placenta (64,000 cells) took 1.35 hours and 4 days, respectively, and this data was not used for the full comparison between LDA and HDP. In terms of memory usage, both LDA and HDP had similar memory consumption in all four datasets.

Table 1. Computational scalability of LDA and HDP model for clustering single-cell RNA-seq data. Table is reproduced from Article I.

	Artificially mixed immune dataset		Pancreas, Tabularis muris		Kidney, Tabularis muris		Decidua/placenta	
	# genes	# cells	# genes	# cells	# genes	# cells	# genes	# cells
	13,000	1,153	23,000	1,961	23,000	2,782	23,000	64,734
LDA	1.7 min/ 2.6 GB		2.8 min/ 4.2 GB		2.3 min/ 6.0 GB		1.35 hrs/208 GB	
HDP	5.7 min/ 2.7 GB		15.2 min/ 4.3 GB		28.1 min/ 6.1 GB		4 days/ 208 GB	

5.1.3 Comparison of LDA Clustering Tools

The Natural Language Processing based Gensim LDA implementation was also compared with scRNA-seq-specific tools such as CELDA [142] and DIMM-SC [141]. Only the first top 2000 highly variable genes were used for DIMM-SC [141], as the execution times extended to several weeks with the full gene lists. Generally, the Gensim defined best clusters with k values based on the lowest BD-index. In the same way, the *DB-index* defined clusters were in accordance with the ones defined by their respective ARI values. This indicated there is a tremendous potential for improving the LDA-based clustering tools that are specifically meant for scRNA-seq analysis. However, Gensim and DIMM-SC showed high variability for repeated experiments. The CELDA [142] results showed less variability for repeated experiments with higher ARI values for the given k as compared to the other two.

Finally, the Bayesian LDA and HDP clusterings were contrasted with the widely used single-cell analysis tool called Seurat, which implements the shared nearest

neighbor (SNN) algorithm for clustering. In terms of intrinsic cluster evaluation metric, the results of Gensim implementation for LDA and HDP clustering showed a comparable result with the Seurat in all three datasets, while Seurat showed consistent results for repeated experiments.

5.2 Contributions of Cell-type Specific Markers in Preeclampsia

Preeclampsia is a pregnancy disorder that is often described based on its onset as early-onset preeclampsia (EOP, before week 34) or late-onset preeclampsia (LOP, after week 34). Previous bulk transcriptome studies [207,208] have identified significant transcriptome changes associated with the disease progression in terms of up regulated genes and down regulated genes in both early and late onset preeclampsia. However, the bulk studies do not allow the detection of the cell-type specific contributions toward the disease progression. The recent single-cell transcriptomics studies on 1st trimester pregnancy [204], menstrual cycle [206], and term pregnancy [205] have opened the door for the study of cellular heterogeneity in the endometrium. The goal of this research project was to integrate the previously published bulk transcriptome data together with recent single-cell transcriptome data in order to identify the cell-type specific contributions in the preeclampsia disease progression.

The over-representation analysis of the cell-type markers showed that the uterine stromal and natural killer cell-type specific marker genes were enriched among the previously detected LOP downregulated ($P = 1.5 \times 10^{-32}$, 5.1-fold) and EOP upregulated genes ($P = 1.3 \times 10^{-22}$, 8.6-fold), respectively. The result indicated the cell-type specific characterization of transcriptomic dynamics in stromal and natural killer cells were associated with EOP and LOP. The over-representation enrichment result for the secretory phase of the menstrual cycle from women that previously had preeclampsia (M-PP) [208] were enriched with perivascular cell markers, but with relatively less striking enrichment p-value.

5.2.1 Contribution of Decidual Stromal Cell Subpopulation Markers in Preeclampsia

The stromal and natural killer cell subpopulation study on the 1st trimester data (E-MTAB-6701) showed three distinct cell subpopulations for both stromal (dS1, dS2 and dS3) (**Figure 9A**) and natural killer cells (dNK1, dNK2, dNK3) (**Figure 10A**) in their differentiation towards the decidualization. dS3 is the decidualized form of the uterine stromal cell identified by the over expression of prolactin as a decidualization marker [204]. The gene over-representation analysis on the three stages of stromal cell differentiation indicated the decidualized stromal cell marker genes had a significant

enrichment towards the downregulated LOP genes (**Figure 9B**). This finding implied LOP affects the normal decidualization of maternal stromal cells during early pregnancy.

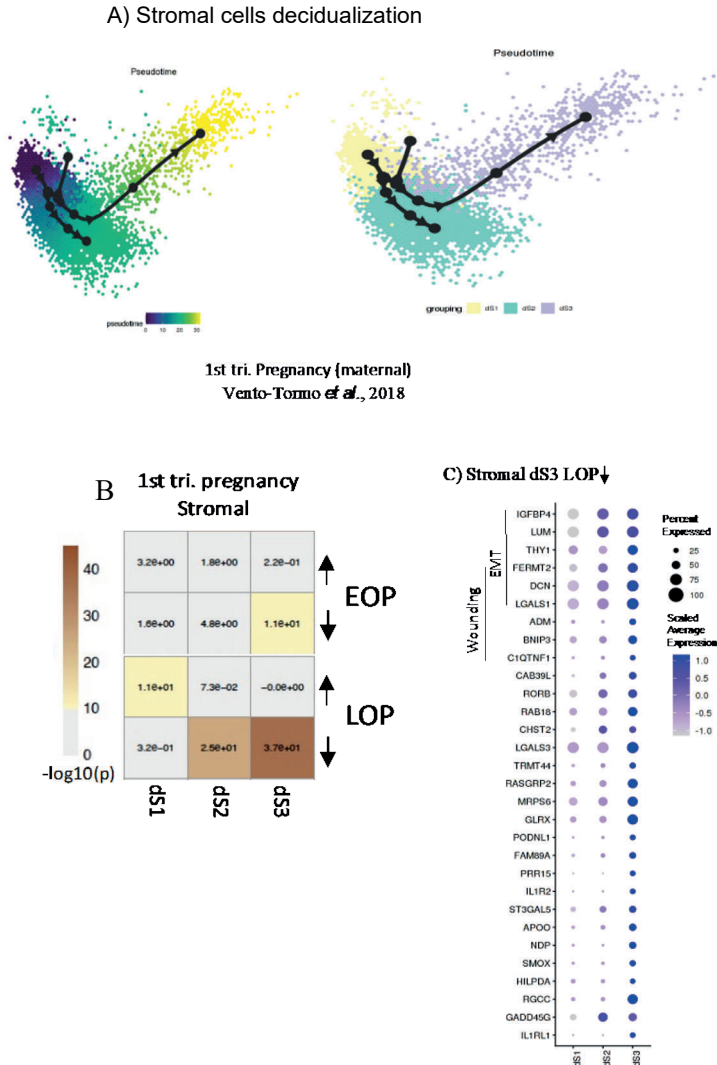


Figure 9. The subpopulations of decidual stromal cells from 1st trimester pregnancy data. (A) The stromal cell decidualization trajectories (B) Gene-set overrepresentation analysis enrichment p-value for the stromal subpopulations specific marker genes on LOP and EOP. The x-axis indicates stromal cell subpopulations, and the y-axis indicates the over-representation enrichment p-value over the negative logarithmic scale on LOP and EOP. (C) Marker genes for the decidualized form of stromal cells that are overrepresented in downregulated genes of LOP with their associated GO terms.

The decidualized stromal cell markers and LOP downregulated genes were associated to GO terms such as “epithelial–mesenchymal transition” (EMT) and

“wounding” (**Figure 9C**). For instance, the expression level of decorin (DCN) and galectin 1 (LGALS1) genes increased over the decidualization process and had their maximum expression in the decidualized cellular state (dS3). The dS1 marker genes were overrepresented among upregulated LOP genes. Similarly, dS3 marker genes were enriched among the downregulated EOP genes, strengthening the hypothesis that the regulatory defect associated to the decidualization of stromal cells significantly contributes to the progression of severe preeclampsia [208,222,223].

5.2.2 Contributions of Decidual Natural Killer Cell Subpopulation Markers in Preeclampsia

The decidual natural killer (dNK) subpopulation (dNK1, dNK2, dNK3) (**Figure 10A**) specific over-representation analysis showed there was no robust cell-type specific enrichment to EOP or LOP. However, the functional enrichment analysis of dNK subpopulation marker genes that were detected to be upregulated in EOP resulted in functional categories including ‘Allograft rejection’ (Hallmark, $P = 4.9 \times 10^{-8}$) and ‘Leukocyte activation’ (GO, $P = 5.1 \times 10^{-8}$). These functional categories are generally linked to reduced maternal immunotolerance. Of these genes, CD3E, TRDC, CORO1A, TMGD2, and ITM2A showed a trend of higher expression in less differentiated dNK subpopulations (dNK1 and dNK2), whereas CD247, CD96, CD7, and CD2 showed a trend of higher expression in more differentiated dNK2 and dNK3 subpopulations (**Figure 10B**).

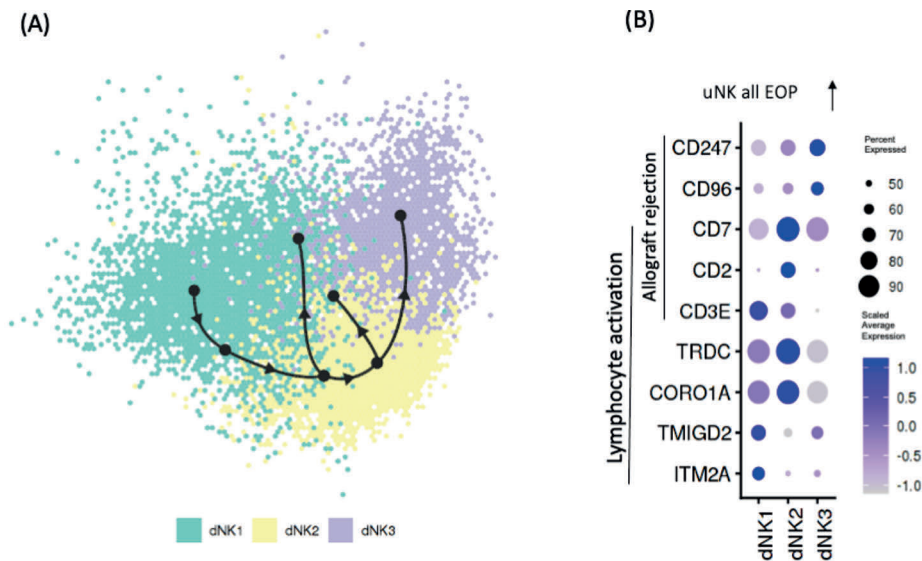


Figure 10. The subpopulation of decidual natural killer cells. (A) The decidualization trajectory for the subpopulations of decidual natural killer cells. (B) the GO terms associated with overrepresented decidual natural killer cell subpopulation among the upregulated EOP genes.

Additionally, perivascular cells had moderate overrepresentation enrichment result over the downregulated genes for samples taken from women that had previously experienced savior preeclampsia [208] with specific GO terms like epithelial-mesenchymal transition ('EMT') ($p\text{-value} = 4.4 \times 10^{-9}$) and 'blood vessel development' ($p\text{-value} = 7.8 \times 10^{-9}$). Interestingly, "epithelial-mesenchymal transition" GO term was also identified in dS3 LOP downregulated signatures supporting previous report from [224] on the association of perivascular and stromal cell contribution towards reproductive disorder.

5.3 Decidualization Regulatory Network Inference for Stromal and Natural Killer Cells

The current knowledge of gene regulatory networks for decidual cell states is established based on bulk *in vivo* transcriptomics studies. Thus, the single-cell resolution transcriptomics analysis enhances the existing cell-type specific decidual gene regulatory network insight. As the two cell types, dS and dNK, are predicted to have contributions to preeclampsia regulation in **Article II**, understanding the mechanism of gene regulatory networks of these cell subpopulations at single-cell resolution is also essential to understand the *in-vivo* uterine states and the decidualization process in more detail and utilize it for therapeutic intervention. Here we studied the transcriptional regulators of dS and dNK subpopulations and predicted the functions of the associated target genes. Additionally, the TF target gene-sets were investigated together with the recent transcriptomic data from pregnancy disorders to predict the translational relevance of the stromal and NK subpopulation-specific regulators.

5.3.1 Subpopulations of Stromal and NK Cells

The 1st trimester decidual stromal cells were previously annotated [204] in three stromal cell subpopulations as dS1, dS2 and dS3 in the order of their differentiation stages. With the re-analysis of stromal cells using the scanpy clustering resolution parameters of 0.3, we identified an additional cluster, which the original author did not identify as a separate cluster (**Figure 11A**), with the upregulated senescent cell marker gene *CXLX8* [225]. However, it is difficult to fully acknowledge the cluster as senescent cells because several other senescent markers were not upregulated. However, NK cell makers such as *GPLY* and *NKG7* were upregulated in this newly identified cluster implying the closer interactions of these cells with the natural killer cells. Hence, this newly identified cluster was annotated as senescent/NK cells. Additionally, the dS1 cell-type clusters were split into two, dS1A and dS1B. dS2 was confirmed with upregulation of pre-decidual marker genes such as *FOXO1* and

LEFTY2. On the other hand, Prolactin (*PRL*) and *IL1B* expression together with other decidual stromal cell (DSC) markers in dS3 identify the cluster as a decidualized cell cluster.

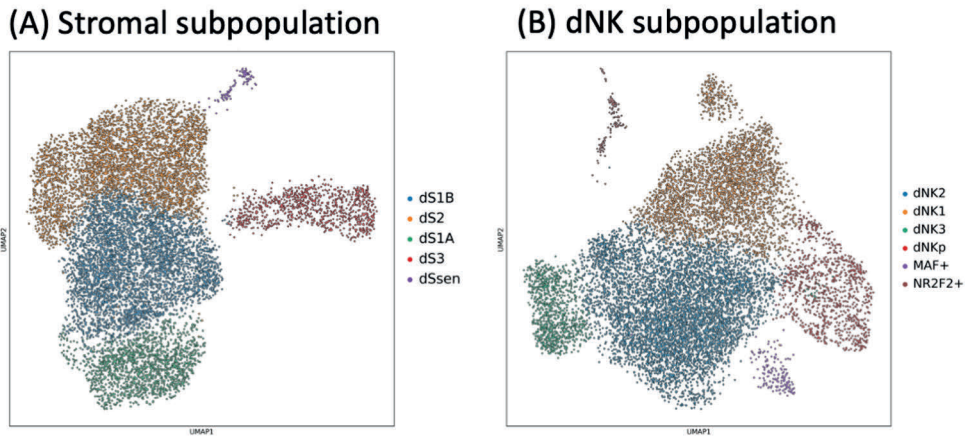


Figure 11. The 2D UMAP visualization of decidual stromal and natural killer cells. (A) The subpopulation of 1st trimester decidual stromal cells with novel cell types identified. (B) Decidual natural killer cell subpopulations with identified potential novel cell-types.

The re-analysis of dNK cells with a cluster resolution parameter of 0.3 showed similarity in the cell-type identity of the clusters with the original cluster annotations by the authors. As a result, the annotations for the dNK cell clusters were adopted from [204] as dNK p, dNK1, dNK2 and dNK3 with their respective cell type markers used as cluster identifiers. However, additional two clusters dNK MAF+ (macrophages) and dNK NRF2F2+ were identified and annotated based on the SCENIC GRN analysis results as explained in the section 4.2.4.

5.3.2 Regulatory Networks of Decidualizing Stromal Cells

The probability of a regulon specificity to each subpopulation was evaluated by the regulon specificity score (RSS) calculated using the Jensen-Shannon distance [226] over the *AUC* value. Further the top 10 regulons based on RSS score that had more than 50 target genes were used for the downstream analysis in each of the subpopulations. As a result, *POLR2A* and *SFR* were identified as undifferentiated fibroblastic-like dS1A regulators, while *FOSL1*, *BHLH40*, *MAFF*, *KLF6*, and *KLF4* were dS1B specific. The “blood vessel morphogenesis” (p -value 1.99×10^{-9}) and “regulation of cell junction assembly” (p -value 2.67×10^{-9}) were the enriched GO terms for dS1A subpopulation specific regulators and their target genes, while the

“hematopoietic or lymphoid organ development” ($p\text{-value } 3.76 \times 10^{-22}$) GO term was for dS1B regulators and their target genes.

The major regulons specific to dS2 subpopulation included *PRDM1*, *NFE2L1*, *FOXP1*, *SOX5*, *STAT2*, *ARID3A*, and *PBX3*. Generally, the top GO terms such as “response to hormone” ($p\text{-value}=1.02 \times 10^{-22}$) and “regulation of Wnt signaling pathway” ($p\text{-value}=3.01 \times 10^{-20}$) were associated with the dS2 specific regulators and their targets. The mouse decidualization TF *PRDM1* was found to regulate genes such as *HAND2*, *LEFTY2*, *WNT5A*, *PRLR* and *IGFBP2* that are involved in the decidualization. The TF *FOXP1* on the other hand was predicted to regulate *LEFTY2* and *PRLR*.

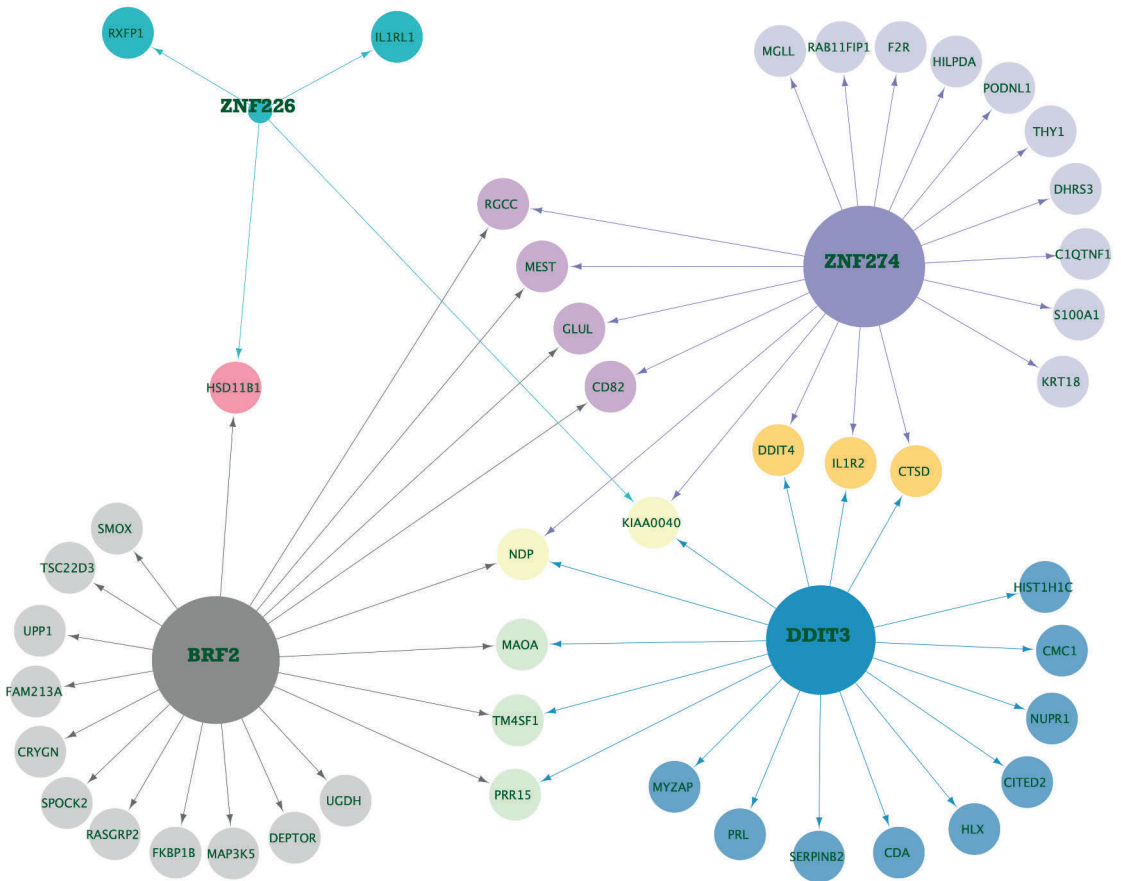


Figure 12. Core gene regulatory networks of decidualized stromal cells (dS3). The arrow in the figure originates from dS3 specific TFs (BRF2, DDIT3, ZNF274, ZNF226) pointing to their target genes. The size of each TF nodes is proportional to the number of target genes it regulates.

The decidualized form of stromal cells (dS3) specific regulators were *BRF2*, *DDIT3*, *ZNF274*, and *ZNF226* (**Figure 12**). The gene expression for *BRF2*, *DDIT3*, and *ZNF226* were specific to dS3 cell type, and these genes were not previously known to be associated with decidualization. The redox-sensing transcription factor *BRF2* regulates secretory gene targets such as *WNT4* [227] that are associated with secretory endometrium. Another dS3-specific TF *DDIT3* targets known decidualization markers such as *PRL*, *IL1R2* and *DDIT4*. These target genes were known to be linked to stress and unfolding protein stress response [228]. The GO terms “regulation of response to endoplasmic reticulum stress” ($p\text{-value}=7.67 \times 10^{-8}$) and “regulation of steroid hormone secretion” ($p\text{-value}=5.54 \times 10^{-7}$) were significantly enriched for the dS3 specific regulons. Additionally, dSsen/nk enriched GO terms “cytolysis” ($p\text{-value}=2.67 \times 10^{-10}$) and positive regulation of chemotaxis ($p\text{-value}=3.84 \times 10^{-7}$) suggested that dSsen/nk cells would appear to be target for immunoclearance by dNK cells.

5.3.3 Regulatory Networks for Decidual NK Cells

Based on the cluster analysis on the decidual natural killer cells, five subpopulations i.e. dNKp, dNK1, dNK2, dNK3 dNK MAF+ and dNK NRF2F2+ were identified (**Figure 11B**). The dNK1 specific regulators were *FOXP2*, *RELB*, *IRX3*, *ZNF100*, and *RREB1* (**Figure 13**), among which *FOXP2*, *RELB*, *IRX3*, and *RREB1* genes had higher expression in the dNK1 cell subpopulation specifically. *RELB* is the known negative regulator of NF κ B pathway targeting anti-inflammatory TF genes such as *NFKBIA* and *STAT3*. It also targets the *CSF1* that promotes the interaction with *ETV* and *SPDL1*. It has been reported that *RELB* genes activates the non-canonical anti-inflammatory state [229] indicating that *RELB* is involved in immunomodulation in NK cells. Both *IRX3* and *RREB1* target the main glycolysis regulators HIFs [204] and another dNK1 regulator *FOXP2* having their target genes GO terms associated with “response to oxygen levels” and “female pregnancy”.

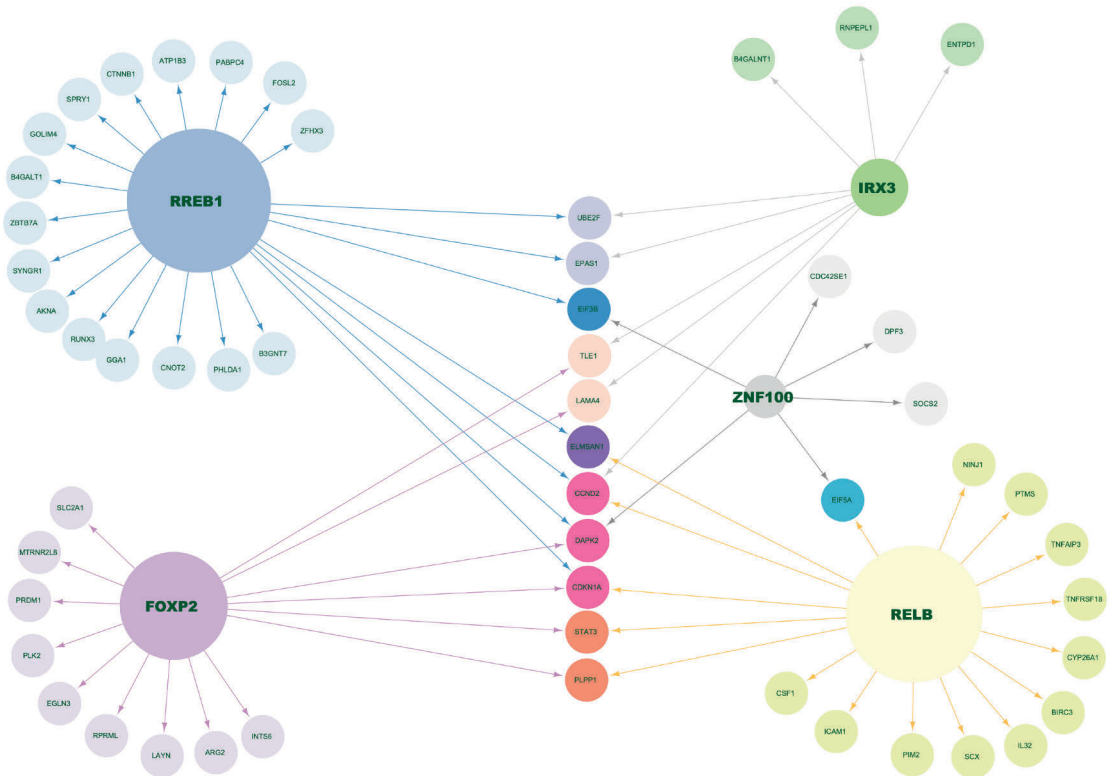


Figure 13. Core gene regulatory networks of immunotolerant natural killer cell subpopulation dNK1. The arrow in the figure originates from dNK1 specific TFs (FOXP2, RREB1, ZNF100, RELB, IRX3) pointing to their target genes. The size of each TF nodes is proportional to the number of target genes it regulates.

Most of the dNK2 identified regulons were shared with the other cell subpopulations. However, *KLF2* and *ZNF143* TFs were found to be dNK2 specific regulators.

The identified dNK3 regulators were *TBX21*(t-Bet), *IRF2* *IRF7*, *TGIF1*, and *FOXP2*. At the gene expression level, all of the genes showed subpopulation-specific high expression patterns. *IRF2* and *TBX21* (T-bet) are known classical regulators of NK development [230]. The target genes for *IRF2* and *IRF7* are mainly involved in interferon response, cytokine regulation and core inflammatory regulators for *STAT1* and *TNF*. *TBX21* (T-bet) had several leukocyte migration related target genes including *TIGIT* whose receptor *PVR* is expressed in *EVT* implying that dNK3 cells interaction with trophoblasts [204]. The functional enrichment analysis for dNK3 regulon targets showed inflammation related terms including response to virus and defense response to symbiont.

5.3.4 Decidual Stromal and NK Cell Regulators in Pregnancy Disorders

The translational aspect of the GRN results for the decidual stromal subpopulation on preeclampsia and recurrent pregnancy loss indicated that dS2 and dS3 specific regulon targets were enriched for the downregulated preeclampsia genes, confirming the previously established claim about the contributions of decidualization defects on preeclampsia. At the same time, the dNK1 regulon target genes were enriched for the disease specific downregulated genes, while dNK3 regulon targets were enriched for downregulated disease genes.

5.4 Single-cell Multi-omics Integrative Data Analysis

Recent advance in the single-cell multi-omics profiling technologies enabled the availability of single-cell multi-omics dataset for integrative analysis. However, the tools and analytical methodologies for the single-cell multi-omics data analysis are not that rampant, even though the bulk multi-omics analysis methodologies and tools were there for quite some time. In this respect, there has been efforts to integrate samples from multiple bulk omic assays using analytical tools such as iClusterBayes [231], intNMF [232], PINSPPlus [233] and CIMLR [234] to unlock the tumour heterogeneity and molecular subtypes of cancer. But the recent technological advances in profiling multi-omics assays at a single-cell resolution widened up the opportunities for investigating cellular heterogeneity at single-cell level. This enabled understanding of biological system in a more detailed way leaving the bioinformatic analysis challenges aside.

5.4.1 Single-cell Multi-omics Data Integration Strategies

The overall workflow starting from the sample extraction, cell disassociation and sorting followed by profiling multiple omics measurements from a single cell gives the single-cell multi-omics dataset (**Figure 14**). The core of the single-cell multi-omics data analysis underlays on identifying cellular heterogeneity from disparate dataset (genomic, transcriptomic, epigenomics, proteomic) profiled from either a single cell or unpaired single-cell multi-omics data. Strategically, the single-cell multi-omics data analysis could be done in three ways: early data integration, late data integration, and intermediate data integration strategies [84,87].

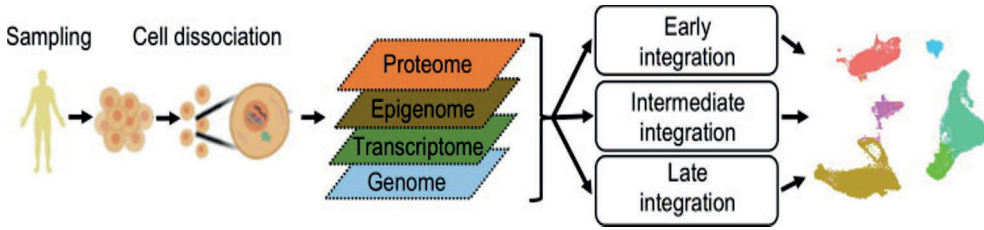


Figure 14. The workflow for single cell multi-omics data analysis. Initially, sample is extracted from tissues or biopsy. Cells are isolated and disassociated before undergoing through single-cell multi-omics profiling protocols. The integrative data analysis of single-cell multi-omics data can be performed in three approaches: early, intermediate and late data integration strategies for improved cell heterogeneity analysis. Figure is reproduced from Article IV.

5.4.2 Early Data Integration Approaches

The early data integration strategy focuses on bringing multiple omics data into one integrated feature matrix to perform the downstream analysis. However, a merged feature matrix brings complicity and increased dimensionality keeping it difficult for direct use. As a result, feature learning and dimensionality reductions over the combined data is essential. Mostly, this approach has been used for integrating multiple single-cell RNA-seq data from multiple sources coupled with normalization and scaling. The major challenge with this approach is the data from different omics layers usually have different feature dimensions and scales, with the potential of the result dominated by the omics layer with more dimensions. Additionally, the sparsity and the high dimensionality of multi-omics data keeps it hard to create a robust common representation across multiple omics data. In this respect, deep AutoEncoder is one of the methods with great potential for applying to single-cell multi-omics analysis that incorporates the data compression for the integrated feature representation.

5.4.3 Late Data Integration Approaches

The late data integration strategy generally applies any of the single-omics analysis among the individual omics data and the data integration takes place at the result level in such a way that the consensus analytical solution is suggested. Particularly, the late integration cluster analysis for single-cell multi-omics data follows two approaches: a two-step and joint-modeling approach. In a two-step approach, first independent cluster analysis is applied for each omics data; then in the second step, the cluster level integration is performed so as to find a common clustering structure representing the multi-omics data. Cluster-of-clusters analysis (COCA) [235], Kernel Learning Integrative Clustering (KLIC) [236], and perturbation-based clustering [237] are methods designed in a two-step late clustering integration

approaches. The joint modeling late cluster integrative approach models the relationship between the local clustering results for finding a robust and improved global clustering solution across multiple omics layers. For instance, SAME-clustering [238] is a method that combines several scRNA-seq data using mixture model ensemble methods in order to create consensus clustering results.

5.4.4 Intermediate Integration Methods

The intermediate approach works in such a way that multiple omics layers are simultaneously used to transform the multi-omics dataset into a single representative data matrix at subspace using similarity-based data integration methods [134,239,240], joint dimensionality reduction [241–244], or statistical modeling [245–248].

5.4.4.1 Similarity-based Methods

The similarity-based data integration methods, such as spectral data integration, utilize pairwise affinity matrix between any pairs of datapoints in the datasets for downstream integrated cluster analysis. Variants of this method are adopted, for instance, in SCHEM [239] and Spectrum [240] for single-cell multi-omics cluster analysis. The similarity-based graph fusion methods integrate graphs from multiple omics layers. In this regard, Seurat4 [134] has implemented weighted-nearest neighbor graph-based integration for single-cell multi-omics cluster analysis.

5.4.4.2 Joint Dimensionality Reduction

The major aim of doing multi-omics cluster analysis is to understand the shared latent structure from multiple high-dimensional datasets to get a comprehensive understanding of the single-cell multi-omics dataset. The recent implementation of manifold alignment-based single-cell multi-omics tools includes MATCHER [241], Manifold-Aligned GAN (MAGAN) [242], Unicom [243], and MMD-MA [244], demonstrating the utility of the manifold alignment algorithms for single-cell multi-omics integrative analysis. There are a few approaches suggested for adopting NMF for multi-omics clustering. The recently introduced tools such as LIGER [249], coupledNMF [250], and MOFA+ [251] demonstrated the utility of NMF algorithms for integrative single-cell multi-omics analysis. Additionally, Seurat3 demonstrated the utility of CCA for the integrative analysis of single-cell RNA-seq and ATAC-seq. A different variant of autoencoders such as variational autoencoders (VAE) has been implemented in totalVI [252] and scMVAE [253] for the integrative analysis of single-cell transcriptomics and proteomics and transcriptomics and open

chromatin accessibility data respectively. Another variant called adversarial autoencoders (AAE) [254] has also been implemented for integrative analysis of single-cell imaging and sequencing data.

5.4.4.3 Model-based Methods

The standard probabilistic Bayesian mixture models are widely used for integrative multi-view cluster analysis in NLP problems. The Bayesian framework for context dependent multi-omics clustering using Dirichlet mixture model was demonstrated by [255]. The first probabilistic cluster assignment takes place at the individual omics layer while extracting global structure that arises from the local cluster assignment using hierarchical Dirichlet mixture models. This shows that a local cluster assignment affects the posterior probabilities of corresponding global cluster assignments. Similarly, another Bayesian frameworks [245,246] were proposed for cluster analysis of multi-omics data in bulk studies. It performs individual omics level clustering separately while simultaneously model the dependencies across individual omics clustering in order to infer a global or consensus clustering solution. Additionally, probabilistic model based algorithm BREM-SC [247], that has been applied in single-cell multi-omics data, utilizes Dirichlet multinomial distribution to model the gene expression and surface protein expression in the CITE-seq data in a framework that introduces specific random effects in order to correlate between different omics. Clonealign [248] also implemented the mean field variational Bayes approach for integrative analysis of the unmatched single-cell gene expression and copy number variation datasets.

5.5 Contributions of the thesis

The studies in **Article I** and **Article IV** focused on the methodological aspect of single-cell RNA-seq cluster analysis and single-cell multi-omics data integration approaches respectively. At the time of the study in **Article I** conceived, there has not been a comprehensive comparative analysis of LDA and HDP model for cluster analysis in different scRNA-seq data. As a result, this study has contributed showcasing the potential use cases, challenges and opportunities of using Dirichlet mixture models for cluster analysis in scRNA-seq and beyond. Similarly, the study in **Article IV** has contributed in suggesting the state-of-art analytical and methodological approaches for integrative single-cell multi-omics analysis. The single-cell RNA-seq being widely used experimental method, there are several methods and pipelines/tools developed for the analysis of scRNA-seq data. Harnessing publicly available tools for scRNA-seq analysis, **Article II** combines three different scRNA-seq data from endometrium and previous bulk RNA-seq

studies to identify cell-type specific marker gene contribution in a disease called preeclampsia. Gaining insight from **Article II** on the decidual stromal and natural killer cells contribution towards preeclampsia, in **Article III**, the gene regulatory networks of decidual a stromal and natural killer subpopulations identified the novel regulators of decidualization for stromal and natural killer cells using the SCENIC single-cell regulatory network analysis pipeline.

6 Discussion

Several studies have been conducted to understand the dynamics of cell heterogeneity using the single-cell omics and multi-omics sequencing. Single-cell RNA-sequencing has been used in different biological and biomedical fields such as immunology [16–22], developmental biology [1–8] and other areas. Other single-cell omics protocols including single-cell epigenetics [25,26], proteomics [34–36] and even single-cell multi-omics are getting growing attention. While the sequencing technologies are advancing with fast phase, the analytical and methodological studies also play a crucial role in interpreting the single-cell sequencing data. Different analytical pipelines or tools have been implemented for single-cell omics and multi-omics analysis.

The Dirichlet mixture models have previously used in other disciplines such as natural language processing while it has rarely been applied to the analysis of single cell cluster analysis. Therefore, **Article I** evaluated the performance and scalability of both parametric and non-parametric Dirichlet mixture models, ie. LDA and HDP, for single-cell RNA-seq cluster analysis using Gensim implementation. Additionally, a comprehensive comparison of these two models on small to medium sized single-cell RNA-seq data was studied. As a result, the relative clustering performance of the LDA and HDP models were evaluated using intrinsic and extrinsic cluster quality metrics and their performance is dataset dependent. This could be attributed to the selection of dataset independent common priori model concentration parameters used in the experiment. However, the optimal clustering results for both models generally approximated the actual biological cell-types. Specially, the nonparametric HDP model is advantageous in that it approximates the number of clusters automatically without the need to predefining the cluster number beforehand. Additionally, HDP could also be used to select optimal model parameters and number of clusters for LDA so that more robots and accurate clustering results were achieved. In this respect, [256] has also demonstrate improved efficiency of LDA models using the effective number of clusters as input parameter from the HDP in the context of text clustering.

Computationally, both models had similar memory consumptions and LDA tend to run faster than HDP. There are few single-cell specific implementation of LDA

and HDP models i.e CELDA [142], DIMM-SC [141] and BISCUIT [88]. However, as inference algorithms used for each of the tools vary, their runtime and memory consumptions difference are attributed to it. In general, Variation inference-based implorations of LDA and HDP (Gensim) run fast and scale for high dimensional datasets as compared to Gibbs Sampling (BISCUIT) and Expectation- maximization (CELDA [142] and DIMM-SC [141]) based implementations.

The second research **Article II** investigated the cell-type specific contributions in pregnancy disorder preeclampsia by integrating the previous bulk transcriptomics with the recent single-cell data. We identified that both decidual stromal and natural killer and their subpopulations specific cells play crucial role in the LOP and EOP. The marker genes for the decidualized form of stromal cell (dS3) were over-represented in downregulated LOP genes suggesting that defects in the decidualization contribute to the disease progression. On the other hand, the genes upregulated in EOP were enriched with dNK markers suggesting a potential overactivation inflammatory type dNK cells during the 1st trimester placentation and spiral artery modeling.

Article III further investigated the gene regulatory network analysis for the subpopulations of stromal and natural killer cells using a healthy 1st trimester pregnancy *in vivo* data. The re-annotation of the clustering on decidual stromal cell resulted in five clusters (dS1A, dS1B, dS2, dS3 and cens/dS3) while the dNK cells were also reannotated as six distinct cell subpopulations (dNK p, dNK1, dNK2, dNK3, MAF+ and NR2F2+). For the Decidualized stromal cell dS3, we identified *BRF2*, *DDIT3*, *ZNF274* and *ZNF226* as their cell-subtype specific regulons. The target genes for these dS3 core regulators included classical decidualization marker such as *PRL* and *WNT4*. Terms related to relative stress tolerance such as oxidative stress response and unfolding protein stress response were detected with the functional enrichment analysis. These terms were in line with previous studies [257–259] associating stress related regulation and decidualization.

The three major subpopulations of decidual natural killer cell have shown distinct cell type specific regulators. The undifferentiated form of decidual natural killer cell dNK1 had specific regulons *FOXP2*, *RELB*, *IRX3*, *ZNF100*, and *RREB1*. The *KLF2* and *ZNF143* TFs were found to be a dNK2 specific regulons while most dNK2 share regulators from both dNK1 and dNK3. dNK1 subpopulations was associated with GOs such as “in utero embryonic development”, “hormone responses”. The dNK3 specific regulons were predicted to be *TBX21* (t-Bet), *IRF2* *IRF7*, *TGIF1* and *FOXN2* with the associated GO terms of “response to virus p-33” and “defense response to symbiont p -25”. Our results further support the view that dNK1 regulators promotes immunotolerance whereas dNK3 regulators such as interferon pathways regulators (IRFs) promoter less immunotolerant state that is also

associated with pregnancy disorders such as preeclampsia and recurrent pregnancy loss [260].

The single-cell multi-omics approaches give the wholistic view of a cell to study cell heterogeneity from multiple omics layer [84,261]. In **Article IV**, the strategies, methodological approaches and tools for the integrative analysis of single-cell multi-omics data were reviewed. Strategically, early, late and intermediate multi-omics integrative approaches were assessed, and most tools recently implemented the intermediate integrative analysis methods, such as spectral, dimensionality reduction and model-based approaches, to identify the cell clusters from more than omics data. However, the late and early integrative approaches had their own merits, for example, the late integrative analysis is advantageous in giving flexibility to apply any single-cell analysis methods independently at each of omics layer and the integrative analysis takes place at a result level. On the other hand, combining the multiple omics data as a single data matrix in an early data integrative analysis approached gives more leverage for feature engineering at the data level reducing the algorithmic or analytical overhead to deal with multiple modalities.

6.1 Limitations

The scope of the study in **Article I** was limited to comparing the application of Dirichlet mixture models, LDA and HDP for cluster analysis in small to medium-sized single-cell RNA-seq data, as the execution time for HDP model clustering takes several days with large datasets. Additionally, multi-parameter tuning for LDA and HDP models is out of the scope of the study. As both models have multiple concentration parameters a priori, the fixed default concentration parameters were used in all experiments. **Article II** combined bulk and single-cell data-derived marker gene over representation analysis identified key cell-type specific contributions for LOP and EOP extending the previous studies on cell-type heterogeneity in EOP and LOP. The study has a limitation on the specificity of time points for each of the samples. For example, preeclampsia samples were collected during delivery for [207] and the transcriptomics landscape might be changing from the initial LOP and EOP during pregnancy [262]. On the other hand, combining single-cell specific cell-type markers for the three different datasets might affect the temporal specificity. Generically, the study gives an insight to conduct further research on cell-type specific contributions for the disease preeclampsia using single-cell technologies. The results in **Article III** predicted the cell sub-type specific core transcriptional regulators. However, these predictions should be validated in future studies for example using knockout strategies.

Article IV reviewed recent implementations of integrative single-cell multi-omics analytical methods and tools using different data integration strategies.

However, the work on comprehensive comparison and benchmarking of such tools are not enough. There is also a limitation in the standardizing the data storage and management. During the time of the manuscript preparation there is no unified consortium that handles the single-cell multi-omics data in a repository despite some efforts taken by Human Cell Atlas project.

6.2 Future directions

Article I suggested the future research in the direction of dataset-specific initialization of model prior concentration parameters and multi-parameter tuning can improve the outcome of clustering results paving the way for an automated cell-type annotation that leads to semi-supervised machine learning analysis. As the single-cell technology improved in the past few years, generalizing numerous high-dimensional data from millions of cells at a time, the analytical aspect of it has to accommodate these growing demands. In this respect, Dirichlet mixture models with efficient inference algorithms that can scale to high dimensional and large datasets had the potential to transform the field of single-cell analysis. The study in **Article II** generically gives an insight to conduct further research on cell-type specific contributions for the disease preeclampsia using single-cell technologies. **Article III** suggest that in the study of endometrium transcriptional regulation studied via GRN analysis can be useful addition to marker gene analysis by providing testable prediction of core regulators for each functionally important subpopulation. In the future, the detected transcription factors can be further studied in cell culture or other models using specific knockouts. Finally, **Article IV** highlights most tools and implementations for single-cell multi-omics data analysis focus on cluster analysis. However, besides the cluster analysis for cell heterogeneity identification, future research works are recommended to focus on enhancing the analytical approach for integrative analysis of networks of gene regulations and motif discovery. Additionally, the single-cell multi-omics data standardization and managements are the areas that must be sought after.

7 Conclusions

Advances in single-cell omics and multi-omics technologies have a tremendous potential to unlock the unknowns of cellular heterogeneity in biological research. In this regard, this thesis showed potential of Bayesian clustering models, ie LDA and HDP, for cell heterogeneity analysis in scRNA-seq data. Additionally, the data integration approaches for single-cell multi-omics data analysis were reviewed. The cell-type specific gene regulatory networks for uterus were also studied using the scRNA-seq sample.

The studies in **Article I** highlighted the importance of Dirichlet mixture models for cellular heterogeneity/cluster analysis in single-cell RNA-seq data. The relative performance of the LDA and HDP models for cluster analysis on small to medium sized datasets was dataset dependent. **Article II** and **Article III** demonstrated how single-cell RNA-seq data analysis and interpretation can help to understand what has not been discovered with previous bulk studies in the human endometrium and the pregnancy disorder preeclampsia. **Article IV** reviewed the strategies and methodologies for integrative single-cell multi-omics data analysis mostly focusing on the cell heterogeneity analysis. However, more work in terms of single-cell multi-omics integrative motif discovery and gene regulatory inference is expected in the future. Additionally, a unified efforts for single-cell multi-omics data storage and management is needed in order to accommodate and use the growing multi-omics data generation for new scientific discovery.

Acknowledgements

I would like to acknowledge all the support and guidance given by principal investigator professor Laura Elo. I would also like to thank Docent Kalle Rytönen for his continued supervision and support to make this PhD path reality. All my co-authors had valuable inputs guiding the course of my research. Dr. Asta Laiho and Dr. Anu Kukkonen-Macchi helped me to integrate into the work culture at Elo Lab.

I would also like to take this opportunity to thank my ENLIGHT-TEN industrial supervisors and managers, Dr. Martin Simonsen, Dr. Leif Schauser and Dr. Vivi Gregersen for their support and guidance during my earlier PhD years. Additionally, I thank all ENLIGHT-TEN PIs, ESRs and coordinators for the amazing courses and experiences.

Finally, while everything seems perfect, there were always latent hands that put courage, curiosity, strength and resilience within me from my childhood. Mr. Philopos Kindo and Mrs. Rebika Gerbo, if this is an achievement, you should take the credit. Mr. Binyam Ayele and Miss. Tihut Ayele, your continuous emotional support brought me here and you deserve an upload. Dr. Bereket Kindo, thank you for your inspirations and being such a role model. Finally, Mr. Eyob Talew and Mr. Bereket Aloto, you have made my stay in Finland smooth and flawless.

April 3, 2023

Nigatu A. Adossa

List of References

1. Tang F, Barbacioru C, Bao S, Lee C, Nordman E, Wang X, et al. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-seq analysis. *Cell Stem Cell*. 2010;6: 468–478. doi:10.1016/j.stem.2010.03.015
2. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods*. 2009;6: 377–382. doi:10.1038/nmeth.1315
3. Yan L, Yang M, Guo H, Yang L, Wu J, Li R, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol*. 2013;20: 1131–1139. doi:10.1038/nsmb.2660
4. Fan X, Zhang X, Wu X, Guo H, Hu Y, Tang F, et al. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol*. 2015;16: 1–17. doi:10.1186/s13059-015-0706-1
5. Sekiguchi R, Martin D, Yamada KM. Single-Cell RNA-seq Identifies Cell Diversity in Embryonic Salivary Glands. *J Dent Res*. 2020;99: 69–78. doi:10.1177/0022034519883888
6. Guo H, Tian L, Zhang JZ, Kitani T, Paik DT, Lee WH, et al. Single-Cell RNA Sequencing of Human Embryonic Stem Cell Differentiation Delineates Adverse Effects of Nicotine on Embryonic Development. *Stem Cell Reports*. 2019;12: 772–786. doi:10.1016/j.stemcr.2019.01.022
7. Shi J, Chen Q, Li X, Zheng X, Zhang Y, Qiao J, et al. Dynamic transcriptional symmetry-breaking in pre-implantation mammalian embryo development revealed by single-cell rna-seq. *Dev*. 2015;142: 3468–3477. doi:10.1242/dev.123950
8. Chu LF, Leng N, Zhang J, Hou Z, Mamott D, Vereide DT, et al. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol*. 2016;17: 173. doi:10.1186/s13059-016-1033-x
9. Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat Commun*. 2017;8: 1–12. doi:10.1038/ncomms15081
10. Kinker GS, Greenwald AC, Tal R, Orlova Z, Cuoco MS, McFarland JM, et al. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat Genet*. 2020;52: 1208–1218. doi:10.1038/s41588-020-00726-6
11. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science (80-)*. 2016;352: 189–196. doi:10.1126/science.aad0501
12. Zhang AW, O’Flanagan C, Chavez EA, Lim JLP, Ceglia N, McPherson A, et al. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat Methods*. 2019;16: 1007–1015. doi:10.1038/s41592-019-0529-1
13. Puram S V., Parikh AS, Tirosh I. Single cell RNA-seq highlights a role for a partial EMT in head and neck cancer. *Mol Cell Oncol*. 2018;5: e1448244. doi:10.1080/23723556.2018.1448244
14. Damiani C, Maspero D, Di Filippo M, Colombo R, Pescini D, Graudenzi A, et al. Integration of single-cell RNA-seq data into population models to characterize cancer metabolism. *PLoS Comput Biol*. 2019;15: e1006733. doi:10.1371/journal.pcbi.1006733

15. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* (80-). 2014;344: 1396–1401. doi:10.1126/science.1254257
16. Chen H, Ye F, Guo G. Revolutionizing immunology with single-cell RNA sequencing. *Cell Mol Immunol*. 2019;16: 242–249. doi:10.1038/s41423-019-0214-4
17. Ranzoni AM, Strzelecka PM, Cvejic A. Application of single-cell RNA sequencing methodologies in understanding haematopoiesis and immunology. *Essays Biochem*. 2019;63: 217–225. doi:10.1042/EBC20180072
18. Neu KE, Tang Q, Wilson PC, Khan AA. Single-Cell Genomics: Approaches and Utility in Immunology. *Trends Immunol*. 2017;38: 140–149. doi:10.1016/j.it.2016.12.001
19. Gomes T, Teichmann SA, Talavera-López C. Immunology Driven by Large-Scale Single-Cell Sequencing. *Trends Immunol*. 2019;40: 1011–1021. doi:10.1016/j.it.2019.09.004
20. Cao Y, Qiu Y, Tu G, Yang C. Single-cell RNA Sequencing in Immunology. *Curr Genomics*. 2020;21: 564–575. doi:10.2174/1389202921999201020203249
21. Eftremova M, Vento-Tormo R, Park JE, Teichmann SA, James KR. Immunology in the Era of Single-Cell Technologies. *Annu Rev Immunol*. 2020;38: 727–757. doi:10.1146/annurev-immunol-090419-020340
22. Zhang J-Y, Wang X-M, Xing X, Xu Z, Zhang C, Song J-W, et al. Single-cell landscape of immunological responses in patients with COVID-19. *Nat Immunol*. 2020;21: 1107–1118.
23. Marioni JC, Arendt D. How single-cell genomics is changing evolutionary and developmental biology. *Annu Rev Cell Dev Biol*. 2017;33: 537–553. doi:10.1146/annurev-cellbio-100616-060818
24. Khrameeva E, Kurochkin I, Han D, Guijarro P, Kanton S, Santel M, et al. Single-cell-resolution transcriptome map of human, chimpanzee, bonobo, and macaque brains. *Genome Res*. 2020;30: 776–789. doi:10.1101/gr.256958.119
25. Xu W, Wen Y, Liang Y, Xu Q, Wang X, Jin W, et al. A plate-based single-cell ATAC-seq workflow for fast and robust profiling of chromatin accessibility. *Nat Protoc*. 2021;16: 4084–4107. doi:10.1038/s41596-021-00583-5
26. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods*. 2014;11: 817–820. doi:10.1038/nmeth.3035
27. Macaulay IC, Voet T. Single Cell Genomics: Advances and Future Perspectives. *PLoS Genet*. 2014;10: e1004126. doi:10.1371/journal.pgen.1004126
28. Trapnell C. Defining cell types and states with single-cell genomics. *Genome Res*. 2015;25: 1491–1498. doi:10.1101/gr.190595.115
29. Paolillo C, Londin E, Fortina P. Single-cell genomics. *Clin Chem*. 2019;65: 972–985.
30. Velmeshev D, Schirmer L, Jung D, Haeussler M, Perez Y, Mayer S, et al. Single-cell genomics identifies cell type-specific molecular changes in autism. *Science* (80-). 2019;364: 685–689. doi:10.1126/science.aav8130
31. Cai X, Evrony GD, Lehmann HS, Elhosary PC, Mehta BK, Poduri A, et al. Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. *Cell Rep*. 2014;8: 1280–1289. doi:10.1016/j.celrep.2014.07.043
32. Lauer S, AVECILLA G, Spealman P, Sethia G, Brandt N, Levy SF, et al. Single-cell copy number variant detection reveals the dynamics and diversity of adaptation. *PLoS Biol*. 2018;16: e3000069. doi:10.1371/journal.pbio.3000069
33. Cheng J, Vanneste E, Konings P, Voet T, Vermeesch JR, Moreau Y. Single-cell copy number variation detection. *Genome Biol*. 2011;12: 1–14. doi:10.1186/gb-2011-12-8-r80
34. Vistain LF, Tay S. Single-cell proteomics. *Trends Biochem Sci*. 2021.
35. Kelly RT. Single-cell proteomics: progress and prospects. *Mol Cell Proteomics*. 2020;19: 1739–1748.

36. Marx V. A dream of single-cell proteomics. *Nat Methods*. 2019;16: 809–812. doi:10.1038/s41592-019-0540-6
37. Auerbach BJ, Hu J, Reilly MP, Li M. Applications of single-cell genomics and computational strategies to study common disease and population-level variation. *Genome Res*. 2021;31: 1728–1741. doi:10.1101/gr.275430.121
38. Single-cell proteomics: challenges and prospects. *Nat Methods*. 2023;20: 317–318. doi:10.1038/s41592-023-01828-9
39. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med*. 2018;50: 1–14. doi:10.1038/s12276-018-0071-8
40. Brehm-Stecher BF, Johnson EA. Single-cell microbiology: tools, technologies, and applications. *Microbiol Mol Biol Rev*. 2004;68: 538–559. doi:10.1128/MMBR.68.3.538-559.2004
41. Julius MH, Masuda T, Herzenberg LA. Demonstration that antigen-binding cells are precursors of antibody-producing cells after purification with a fluorescence-activated cell sorter. *Proc Natl Acad Sci U S A*. 1972;69: 1934–1938. doi:10.1073/pnas.69.7.1934
42. Nichterwitz S, Chen G, Aguila Benitez J, Yilmaz M, Storvall H, Cao M, et al. Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. *Nat Commun*. 2016;7: 12139. doi:10.1038/ncomms12139
43. Marcus JS, Anderson WF, Quake SR. Microfluidic single-cell mRNA isolation and analysis. *Anal Chem*. 2006;78: 3084–3089. doi:10.1021/ac0519460
44. Guo F, Li L, Li J, Wu X, Hu B, Zhu P, et al. Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res*. 2017;27: 967–988. doi:10.1038/cr.2017.82
45. Kashima Y, Sakamoto Y, Kaneko K, Seki M, Suzuki Y, Suzuki A. Single-cell sequencing techniques from individual to multiomics analyses. *Experimental and Molecular Medicine*. 2020. pp. 1419–1427. doi:10.1038/s12276-020-00499-2
46. Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods*. 2017;14: 955–958. doi:10.1038/nmeth.4407
47. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*. 2015;161: 1202–1214. doi:10.1016/j.cell.2015.05.002
48. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*. 2015;161: 1187–1201. doi:10.1016/j.cell.2015.04.044
49. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. 2017;8: 14049. doi:10.1038/ncomms14049
50. Hashimoto S, Tabuchi Y, Yurino H, Hirohashi Y, Deshimaru S, Asano T, et al. Comprehensive single-cell transcriptome analysis reveals heterogeneity in endometrioid adenocarcinoma tissues. *Sci Rep*. 2017;7: 14225. doi:10.1038/s41598-017-14676-3
51. Gierahn TM, Wadsworth MH, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-Well: Portable, low-cost rna sequencing of single cells at high throughput. *Nat Methods*. 2017;14: 395–398. doi:10.1038/nmeth.4179
52. Spits C, Le Caignec C, De Rycke M, Van Haute L, Van Steirteghem A, Liebaers I, et al. Whole-genome multiple displacement amplification from single cells. *Nat Protoc*. 2006;1: 1965–1970. doi:10.1038/nprot.2006.326
53. Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science (80-)*. 2012;338: 1622–1626. doi:10.1126/science.1229164
54. Telenius H, Carter NP, Bebb CE, Nordenskjöld M, Ponder BAJ, Tunnacliffe A. Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics*. 1992;13: 718–725. doi:10.1016/0888-7543(92)90147-K

55. Gu H, Raman AT, Wang X, Gaiti F, Chaligne R, Mohammad AW, et al. Smart-RRBS for single-cell methylome and transcriptome analysis. *Nat Protoc.* 2021;16: 4004–4030. doi:10.1038/s41596-021-00571-9
56. Buenrostro JD, Wu B, Litzemberger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature.* 2015;523: 486–490. doi:10.1038/nature14590
57. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature.* 2013;502: 59–64. doi:10.1038/nature12593
58. Liu Y, Singh AK. Microfluidic Platforms for Single-Cell Protein Analysis. *J Lab Autom.* 2013;18: 446–454. doi:10.1177/2211068213494389
59. Cheung RK, Utz PJ. Screening: CyTOF—the next generation of cell detection. *Nat Rev Rheumatol.* 2011;7: 502–503. doi:10.1038/nrrheum.2011.110
60. Hagemann-Jensen M, Ziegenhain C, Chen P, Ramsköld D, Hendriks GJ, Larsson AJM, et al. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat Biotechnol.* 2020;38: 708–714. doi:10.1038/s41587-020-0497-0
61. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc.* 2014;9: 171–181. doi:10.1038/nprot.2014.006
62. Ramsköld D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol.* 2012;30: 777–782. doi:10.1038/nbt.2282
63. Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, et al. Quartz-Seq: A highly reproducible and sensitive single-cell RNA sequencing method, reveals nongenetic gene-expression heterogeneity. *Genome Biol.* 2013;14: 3097. doi:10.1186/gb-2013-14-4-r31
64. Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol.* 2015;33: 1165–1172. doi:10.1038/nbt.3383
65. Minakshi P, Kumar R, Ghosh M, Saini HM, Ranjan K, Brar B, et al. Single-cell proteomics: Technology and applications. In: Barh D, Azevedo VBT-S-CO, editors. *Single-Cell Omics: Volume 1: Technological Advances and Applications.* Academic Press; 2019. pp. 283–318. doi:10.1016/B978-0-12-814919-5.00014-2
66. Perez OD, Nolan GP. Simultaneous measurement of multiple active kinase states using polychromatic flow cytometry. *Nat Biotechnol.* 2002;20: 155–162. doi:10.1038/nbt0202-155
67. Su Y, Shi Q, Wei W. Single cell proteomics in biomedicine: High-dimensional data acquisition, visualization, and analysis. *Proteomics.* 2017. doi:10.1002/pmic.201600267
68. Hughes AJ, Spelke DP, Xu Z, Kang CC, Schaffer D V., Herr AE. Single-cell western blotting. *Nat Methods.* 2014. doi:10.1038/nmeth.2992
69. Budnik B, Levy E, Harmange G, Slavov N. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol.* 2018;19. doi:10.1186/s13059-018-1547-5
70. Tajik M, Baharfar M, Donald WA. Single-cell mass spectrometry. *Trends Biotechnol.* 2022;40: 1374–1392. doi:10.1016/j.tibtech.2022.04.004
71. Hu Y, An Q, Sheu K, Trejo B, Fan S, Guo Y. Single cell multi-omics technology: Methodology and application. *Front Cell Dev Biol.* 2018;6. doi:10.3389/fcell.2018.00028
72. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods.* 2015;12: 519–522. doi:10.1038/nmeth.3370
73. Dey SS, Kester L, Spanjaard B, Bienko M, Van Oudenaarden A. Integrated genome and transcriptome sequencing of the same cell. *Nat Biotechnol.* 2015;33: 285–289. doi:10.1038/nbt.3129

74. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods*. 2016;13: 229–232. doi:10.1038/nmeth.3728
75. Hu Y, Huang K, An Q, Du G, Hu G, Xue J, et al. Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol*. 2016;17. doi:10.1186/s13059-016-0950-z
76. Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res*. 2016;26: 304–319. doi:10.1038/cr.2016.23
77. Pott S. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *Elife*. 2017;6. doi:10.7554/eLife.23203
78. Clark SJ, Argelaguet R, Kapourani CA, Stubbs TM, Lee HJ, Alda-Catalinas C, et al. ScNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells e. *Nat Commun*. 2018;9. doi:10.1038/s41467-018-03149-4
79. Liu L, Liu C, Quintero A, Wu L, Yuan Y, Wang M, et al. Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nat Commun*. 2019;10. doi:10.1038/s41467-018-08205-7
80. Darmanis S, Gallant CJ, Marinescu VD, Niklasson M, Segerman A, Flamourakis G, et al. Simultaneous Multiplexed Measurement of RNA and Proteins in Single Cells. *Cell Rep*. 2016;14: 380–389. doi:10.1016/j.celrep.2015.12.021
81. Macaulay IC, Ponting CP, Voet T. Single-Cell Multiomics: Multiple Measurements from Single Cells. *Trends in Genetics*. 2017. pp. 155–168. doi:10.1016/j.tig.2016.12.003
82. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, et al. Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotechnol*. 2017;35: 936–939. doi:10.1038/nbt.3973
83. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods*. 2017;14: 865–868. doi:10.1038/nmeth.4380
84. Lee J, Hyeon DY, Hwang D. Single-cell multiomics: technologies and data analysis methods. *Exp Mol Med*. 2020;52: 1428–1442. doi:10.1038/s12276-020-0420-2
85. Frei AP, Bava FA, Zunder ER, Hsieh EWY, Chen SY, Nolan GP, et al. Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat Methods*. 2016;13: 269–275. doi:10.1038/nmeth.3742
86. Gerlach JP, van Buggenum JAG, Tanis SEJ, Hogeweg M, Heuts BMH, Muraro MJ, et al. Combined quantification of intracellular (phospho-)proteins and transcriptomics from fixed single cells. *Sci Rep*. 2019;9: 1469. doi:10.1038/s41598-018-37977-7
87. Bock C, Farlik M, Sheffield NC. Multi-Omics of Single Cells: Strategies and Applications. *Trends in Biotechnology*. 2016. pp. 605–608. doi:10.1016/j.tibtech.2016.04.004
88. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopaeki C, Prabhakaran S, et al. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell*. 2018;174: 1293–1308.e36. doi:10.1016/j.cell.2018.05.060
89. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience*. 2018;7. doi:10.1093/gigascience/giy059
90. Lareau CA, Ma S, Duarte FM, Buenrostro JD. Inference and effects of barcode multiplets in droplet-based single-cell assays. *Nat Commun*. 2020;11: 866. doi:10.1038/s41467-020-14667-5
91. Wolock SL, Lopez R, Klein AM. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst*. 2019;8: 281–291.e9. doi:10.1016/j.cels.2018.11.005
92. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst*. 2019;8: 329–337.e4. doi:10.1016/j.cels.2019.03.003

93. Lun ATL, Riesenfeld S, Andrews T, Dao TP, Gomes T, Marioni JC. EmptyDrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* 2019;20: 63. doi:10.1186/s13059-019-1662-y
94. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, et al. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* 2016;17: 29. doi:10.1186/s13059-016-0888-1
95. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* 2016;17: 75. doi:10.1186/s13059-016-0947-7
96. Bacher R, Chu LF, Leng N, Gasch AP, Thomson JA, Stewart RM, et al. SCnorm: Robust normalization of single-cell RNA-seq data. *Nat Methods.* 2017;14: 584–586. doi:10.1038/nmeth.4263
97. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* 2019;20: 296. doi:10.1186/s13059-019-1874-1
98. Tang W, Bertaux F, Thomas P, Stefanelli C, Saint M, Marguerat S, et al. BayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data. *Bioinformatics.* 2020;36: 1174–1181. doi:10.1093/bioinformatics/btz726
99. Andrews TS, Kiselev VY, McCarthy D, Hemberg M. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nat Protoc.* 2021;16: 1–9. doi:10.1038/s41596-020-00409-w
100. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol.* 2018;36: 421–427. doi:10.1038/nbt.4091
101. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol.* 2018;36: 411–420. doi:10.1038/nbt.4096
102. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive Integration of Single-Cell Data. *Cell.* 2019;177: 1888–1902.e21. doi:10.1016/j.cell.2019.05.031
103. Shaham U, Stanton KP, Zhao J, Li H, Raddassi K, Montgomery R, et al. Removal of batch effects using distribution-matching residual networks. *Bioinformatics.* 2017;33: 2539–2546. doi:10.1093/bioinformatics/btx196
104. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods.* 2019;16: 1289–1296. doi:10.1038/s41592-019-0619-0
105. Hie B, Brynora B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat Biotechnol.* 2019;37: 685–691. doi:10.1038/s41587-019-0113-3
106. Polański K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park JE. BBKNN: Fast batch alignment of single cell transcriptomes. *Bioinformatics.* 2020;36: 964–965. doi:10.1093/bioinformatics/btz625
107. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. *Nat Methods.* 2019;16: 715–721. doi:10.1038/s41592-019-0494-8
108. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007;8: 118–127. doi:10.1093/biostatistics/kxj037
109. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell.* 2019;177: 1873–1887.e17. doi:10.1016/j.cell.2019.05.006
110. Lin Y, Ghazanfar S, Wang KYX, Gagnon-Bartsch JA, Lo KK, Su X, et al. ScMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proc Natl Acad Sci U S A.* 2019;116: 9775–9784. doi:10.1073/pnas.1820006116

111. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert JP. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun.* 2018;9. doi:10.1038/s41467-017-02554-5
112. Chen M, Zhou X. Controlling for Confounding Effects in Single Cell RNA Sequencing Studies Using both Control and Target Genes. *Sci Rep.* 2017;7: 13587. doi:10.1038/s41598-017-13665-w
113. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol.* 2014;32: 896–902. doi:10.1038/nbt.2931
114. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol.* 2015;33: 155–160. doi:10.1038/nbt.3102
115. Scialdone A, Natarajan KN, Saraiva LR, Proserpio V, Teichmann SA, Stegle O, et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods.* 2015;85: 54–61. doi:10.1016/j.ymeth.2015.06.021
116. van Dijk D, Sharma R, Nainys J, Yin K, Kathail P, Carr AJ, et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell.* 2018;174: 716-729.e27. doi:10.1016/j.cell.2018.05.061
117. Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, et al. SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods.* 2018;15: 539–542. doi:10.1038/s41592-018-0033-z
118. Chen M, Zhou X. VIPER: Variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. *Genome Biol.* 2018;19: 196. doi:10.1186/s13059-018-1575-1
119. Gong W, Kwak IY, Pota P, Koyano-Nakagawa N, Garry DJ. DrImpute: Imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics.* 2018;19: 220. doi:10.1186/s12859-018-2226-y
120. Wang J, Agarwal D, Huang M, Hu G, Zhou Z, Ye C, et al. Data denoising with transfer learning in single-cell transcriptomics. *Nat Methods.* 2019;16: 875–878. doi:10.1038/s41592-019-0537-1
121. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun.* 2019;10: 390. doi:10.1038/s41467-018-07931-2
122. Arisdakessian C, Poirion O, Yunits B, Zhu X, Garmire LX. DeepImpute: An accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. *Genome Biol.* 2019;20: 211. doi:10.1186/s13059-019-1837-6
123. Peng T, Zhu Q, Yin P, Tan K. SCRABBLE: Single-cell RNA-seq imputation constrained by bulk RNA-seq data. *Genome Biol.* 2019;20: 88. doi:10.1186/s13059-019-1681-8
124. Hou W, Ji Z, Ji H, Hicks SC. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol.* 2020;21: 218. doi:10.1186/s13059-020-02132-x
125. Ly LH, Vingron M. Effect of imputation on gene network reconstruction from single-cell RNA-seq data. *Patterns.* 2022;3: 100414. doi:10.1016/j.patter.2021.100414
126. Svensson V, Natarajan KN, Ly LH, Miragaia RJ, Labalette C, Macaulay IC, et al. Power analysis of single-cell rna-sequencing experiments. *Nat Methods.* 2017;14: 381–387. doi:10.1038/nmeth.4220
127. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell.* 2017;65: 631-643.e4. doi:10.1016/j.molcel.2017.01.023
128. Grün D, Kester L, Van Oudenaarden A. Validation of noise models for single-cell transcriptomics. *Nat Methods.* 2014;11: 637–640. doi:10.1038/nmeth.2930
129. Andrews TS, Hemberg M. M3Drop: Dropout-based feature selection for scRNASeq. *Bioinformatics.* 2019;35: 2865–2867. doi:10.1093/bioinformatics/bty1044
130. Jiang L, Chen H, Pinello L, Yuan GC. GiniClust: Detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol.* 2016;17. doi:10.1186/s13059-016-1010-4

131. Van Der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res.* 2008;9: 2579–2625.
132. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. *J Open Source Softw.* 2018;3: 861. doi:10.21105/joss.00861
133. Kiselev VY, Kirschnner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: Consensus clustering of single-cell RNA-seq data. *Nat Methods.* 2017;14: 483–486. doi:10.1038/nmeth.4236
134. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. *Cell.* 2021;184: 3573–3587.e29. doi:10.1016/j.cell.2021.04.048
135. Zhu X, Ching T, Pan X, Weissman SM, Garmire L. Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. *PeerJ.* 2017;2017: e2888–e2888. doi:10.7717/peerj.2888
136. Zhang S, Yang L, Yang J, Lin Z, Ng MK. Dimensionality reduction for single cell RNA sequencing data using constrained robust non-negative matrix factorization. *NAR Genomics Bioinforma.* 2020;2. doi:10.1093/nargab/lqaa064
137. Woo J, Winterhoff BJ, Starr TK, Aliferis C, Wang J. De novo prediction of cell-type complexity in single-cell RNA-seq and tumor microenvironments. *Life Sci alliance.* 2019;2: e201900443. doi:10.26508/lsa.201900443
138. Kotliar D, Veres A, Nagy MA, Tabrizi S, Hodis E, Melton DA, et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife.* 2019;8: e43803. doi:10.7554/eLife.43803
139. Zhu L, Lei J, Klei L, Devlin B, Roeder K. Semisoft clustering of single-cell data. *Proc Natl Acad Sci U S A.* 2018/12/26. 2019;116: 466–471. doi:10.1073/pnas.1817715116
140. Blei DM, Ng AY, Jordan MI, Koltcov S, Koltsova O, Nikolenko S, et al. Latent Dirichlet allocation. *J Mach Learn Res.* 2003;3: 993–1022. doi:10.1016/b978-0-12-411519-4.00006-9
141. Sun Z, Wang T, Deng K, Wang XF, Lafyatis R, Ding Y, et al. DIMM-SC: A Dirichlet mixture model for clustering droplet-based single cell transcriptomic data. Sahinalp C, editor. *Bioinformatics.* 2018;34: 139–146. doi:10.1093/bioinformatics/btx490
142. Wang Z, Yang S, Koga Y, Corbett SE, Johnson WE, Yajima M, et al. Celda: A Bayesian model to perform co-clustering of genes into modules and cells into subpopulations using single-cell RNA-seq data. *bioRxiv.* 2021; 2020.11.16.373274. doi:10.1101/2020.11.16.373274
143. duVerle DA, Yotsukura S, Nomura S, Aburatani H, Tsuda K. CellTree: An R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data. *BMC Bioinformatics.* 2016;17: 363. doi:10.1186/s12859-016-1175-6
144. Shin J, Berg DA, Zhu Y, Shin JY, Song J, Bonaguidi MA, et al. Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. *Cell Stem Cell.* 2015;17: 360–372. doi:10.1016/j.stem.2015.07.013
145. Ji Z, Ji H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* 2016;44: e117. doi:10.1093/nar/gkw430
146. Coifman RR, Lafon S. Diffusion maps. *Appl Comput Harmon Anal.* 2006;21: 5–30. doi:10.1016/j.acha.2006.04.006
147. Angerer P, Haghverdi L, Büttner M, Theis FJ, Marr C, Buettner F. Destiny: Diffusion maps for large-scale single-cell data in R. *Bioinformatics.* 2016;32: 1241–1243. doi:10.1093/bioinformatics/btv715
148. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol.* 2016;34: 637–645. doi:10.1038/nbt.3569
149. Saul L, Roweis S. An introduction to locally linear embedding. Unpubl Available <http://www.cs.toronto.edu/~roweis/lle/papers/lleintro.pdf> 2000;7: 1–13. Available: <https://www.cs.nyu.edu/~roweis/lle/papers/lleintro.pdf>
150. Welch JD, Hartemink AJ, Prins JF. SLICER: Inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biol.* 2016;17: 106. doi:10.1186/s13059-016-0975-3

151. Way GP, Greene CS. Bayesian deep learning for single-cell analysis. *Nat Methods*. 2018;15: 1009–1010. doi:10.1038/s41592-018-0230-9
152. Rashid S, Shah S, Bar-Joseph Z, Pandya R. Dhaka: variational autoencoder for unmasking tumor heterogeneity from single cell genomic data. *Bioinformatics*. 2021;37: 1535–1543. doi:10.1093/bioinformatics/btz095
153. Deng Y, Bao F, Dai Q, Wu LF, Altschuler SJ. Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. *Nat Methods*. 2019;16: 311–314. doi:10.1038/s41592-019-0353-7
154. Wang D, Gu J. VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by Deep Variational Autoencoder. *Genomics, Proteomics Bioinforma*. 2018;16: 320–331. doi:10.1016/j.gpb.2018.08.003
155. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*. 2014;32: 381–386. doi:10.1038/nbt.2859
156. Suomi T, Seyednasrullah F, Jaakkola MK, Faux T, Elo LL. ROTS: An R package for reproducibility-optimized statistical testing. *PLoS Comput Biol*. 2017;13. doi:10.1371/journal.pcbi.1005562
157. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11: R106. doi:10.1186/gb-2010-11-10-r106
158. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2009;26: 139–140. doi:10.1093/bioinformatics/btp616
159. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet*. 2015;16: 133–145. doi:10.1038/nrg3833
160. Bacher R, Kendzierski C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology*. 2016. doi:10.1186/s13059-016-0927-y
161. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. *PLoS Comput Biol*. 2015;11. doi:10.1371/journal.pcbi.1004575
162. Katayama S, Töhönen V, Linnarsson S, Kere J. SAMstr: Statistical test for differential expression in single-cell transcriptome with spike-in normalization. *Bioinformatics*. 2013;29: 2943–2945. doi:10.1093/bioinformatics/btt511
163. Miao Z, Deng K, Wang X, Zhang X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics*. 2018;34: 3223–3224. doi:10.1093/bioinformatics/bty332
164. Wang T, Nabavi S. SigEMD: A powerful method for differential gene expression analysis in single-cell RNA sequencing data. *Methods*. 2018;145: 25–32. doi:10.1016/j.ymeth.2018.04.017
165. Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics*. 2019;20: 40. doi:10.1186/s12859-019-2599-6
166. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol*. 2019;15: e8746. doi:10.15252/msb.20188746
167. Saliba AE, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: Advances and future challenges. *Nucleic Acids Res*. 2014;42: 8845–8860. doi:10.1093/nar/gku555
168. Kharchenko P V., Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. *Nat Methods*. 2014;11: 740–742. doi:10.1038/nmeth.2967
169. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol*. 2015;16. doi:10.1186/s13059-015-0844-5
170. Korthauer KD, Chu LF, Newton MA, Li Y, Thomson J, Stewart R, et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol*. 2016;17: 222. doi:10.1186/s13059-016-1077-y

171. Nabavi S, Schmolze D, Maitituoheti M, Malladi S, Beck AH. EMDomics: A robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. *Bioinformatics*. 2016;32: 533–541. doi:10.1093/bioinformatics/btv634
172. Delmans M, Hemberg M. Discrete distributional differential expression (D3E) - a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics*. 2016;17. doi:10.1186/s12859-016-0944-6
173. Vu TN, Wills QF, Kalari KR, Niu N, Wang L, Rantalainen M, et al. Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics*. 2016. doi:10.1093/bioinformatics/btw202
174. Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. *Nat Methods*. 2017;14: 309–315. doi:10.1038/nmeth.4150
175. Sonesson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods*. 2018;15: 255–261. doi:10.1038/nmeth.4612
176. Van den Berge K, Roux de Bézieux H, Street K, Saelens W, Cannoodt R, Saeys Y, et al. Trajectory-based differential expression analysis for single-cell sequencing data. *Nat Commun*. 2020;11: 1201. doi:10.1038/s41467-020-14766-3
177. Nueda MJ, Tarazona S, Conesa A. Next maSigPro: Updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics*. 2014;30: 2598–2602. doi:10.1093/bioinformatics/btu333
178. Sanavia T, Finotello F, Di Camillo B. FunPat: Function-based pattern analysis on RNA-seq time series data. *BMC Genomics*. 2015;16: S2. doi:10.1186/1471-2164-16-S6-S2
179. Bacher R, Leng N, Chu LF, Ni Z, Thomson JA, Kendziorski C, et al. Trendy: Segmented regression analysis of expression dynamics in high-throughput ordered profiling experiments. *BMC Bioinformatics*. 2018;19: 380. doi:10.1186/s12859-018-2405-x
180. Lönnberg T, Svensson V, James KR, Fernandez-Ruiz D, Sebina I, Montandon R, et al. Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves TH1/TFH fate bifurcation in malaria. *Sci Immunol*. 2017;2. doi:10.1126/sciimmunol.aal2192
181. Nelder, J. A., & Wedderburn RWM. *Generalized Linear Models Why Generalized Linear Models?* Journal of the Royal Statistical Society. Routledge; 1972.
182. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: Single-cell regulatory network inference and clustering. *Nat Methods*. 2017;14: 1083–1086. doi:10.1038/nmeth.4463
183. Liu H, Li P, Zhu M, Wang X, Lu J, Yu T. Nonlinear network reconstruction from gene expression data using marginal dependencies measured by DCOL. *PLoS One*. 2016;11: e0158247. doi:10.1371/journal.pone.0158247
184. Woodhouse S, Piterman N, Wintersteiger CM, Göttgens B, Fisher J. SCNS: A graphical tool for reconstructing executable regulatory networks from single-cell genomic data. *BMC Syst Biol*. 2018;12: 59. doi:10.1186/s12918-018-0581-y
185. Lim CY, Wang H, Woodhouse S, Piterman N, Wernisch L, Fisher J, et al. BTR: Training asynchronous Boolean models using single-cell expression data. *BMC Bioinformatics*. 2016;17: 355. doi:10.1186/s12859-016-1235-y
186. Hamey FK, Nestorowa S, Kinston SJ, Kent DG, Wilson NK, Gottgens B. Reconstructing blood stem cell regulatory network models from single-cell molecular profiles. *Proc Natl Acad Sci U S A*. 2017;114: 5822–5829. doi:10.1073/pnas.1610609114
187. Chen S, Mar JC. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*. 2018;19: 232. doi:10.1186/s12859-018-2217-z
188. Ocone A, Haghverdi L, Mueller NS, Theis FJ. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics*. 2015;31: i89–i96. doi:10.1093/bioinformatics/btv257

189. Matsumoto H, Kiryu H, Furusawa C, Ko MSH, Ko SBH, Gouda N, et al. SCODE: An efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics*. 2017;33: 2314–2321. doi:10.1093/bioinformatics/btx194
190. Matsumoto H, Kiryu H. SCoup: Probabilistic model based on the Ornstein-Uhlenbeck process to analyze single-cell expression data during differentiation. *BMC Bioinformatics*. 2016;17: 232. doi:10.1186/s12859-016-1109-3
191. Nguyen H, Tran D, Tran B, Pehlivan B, Nguyen T. A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. *Brief Bioinform*. 2021;22. doi:10.1093/bib/bbaa190
192. Hardoon DR, Shawe-Taylor J. Sparse canonical correlation analysis. *Mach Learn*. 2011;83: 331–353. doi:10.1007/s10994-010-5222-7
193. Parkhomenko E, Tritchler D, Beyene J. Sparse canonical correlation analysis with application to genomic data integration. *Stat Appl Genet Mol Biol*. 2009;8. doi:10.2202/1544-6115.1406
194. Blei DM, Edu BB, Ng AY, Edu AS, Jordan MI, Edu JB. 10.1162/jmlr.2003.3.4-5.993. CrossRef List Deleted DOIs. 2000;1: 993–1022. doi:10.1162/jmlr.2003.3.4-5.993
195. Koltcov S, Koltsova O, Nikolenko S. Latent dirichlet allocation. Proceedings of the 2014 ACM conference on Web science - WebSci '14. 2014. doi:10.1145/2615569.2615680
196. Dey KK, Hsiao CJ, Stephens M. Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genet*. 2017;13. doi:10.1371/journal.pgen.1006599
197. Prabhakaran S, Azizi E, Carr A, Pe'er D. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. 33rd International Conference on Machine Learning, ICML 2016. International Machine Learning Society (IMLS); 2016. pp. 1691–1715.
198. Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet processes. *J Am Stat Assoc*. 2006;101: 1566–1581. doi:10.1198/016214506000000302
199. Gill J. Introduction to Applied Bayesian Statistics and Estimation for Social Scientists. *J Am Stat Assoc*. 2008;103: 1322–1323. doi:10.1198/jasa.2008.s250
200. Fan W, Bouguila N. Online variational learning of generalized Dirichlet mixture models with feature selection. *Neurocomputing*. 2014. doi:10.1016/j.neucom.2012.09.047
201. Blei DM, Kucukelbir A, McAuliffe JD. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*. 2017. pp. 859–877. doi:10.1080/01621459.2017.1285773
202. Bravo González-Blas C, Minnoye L, Papasokrati D, Aibar S, Hulselmans G, Christiaens V, et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat Methods*. 2019;16: 397–400. doi:10.1038/s41592-019-0367-1
203. Wang L, Wang X. Hierarchical Dirichlet process model for gene expression clustering Computational methods for biomarker discovery and systems biology research. *Eurasip J Bioinforma Syst Biol*. 2013;2013. doi:10.1186/1687-4153-2013-5
204. Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, et al. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature*. 2018;563: 347–353. doi:10.1038/s41586-018-0698-6
205. Pique-Regi R, Romero R, Tarca AL, Sandler ED, Xu Y, Garcia-Flores V, et al. Single cell transcriptional signatures of the human placenta in term and preterm parturition. *Elife*. 2019;8. doi:10.7554/eLife.52004
206. Wang W, Vilella F, Alama P, Moreno I, Mignardi M, Isakova A, et al. Single-cell transcriptomic atlas of the human endometrium during the menstrual cycle. *Nat Med*. 2020;26: 1644–1653. doi:10.1038/s41591-020-1040-z
207. Tong J, Zhao W, Lv H, Li WP, Chen ZJ, Zhang C. Transcriptomic Profiling in Human Decidua of Severe Preeclampsia Detected by RNA Sequencing. *J Cell Biochem*. 2018;119: 607–615. doi:10.1002/jcb.26221

208. Garrido-Gomez T, Castillo-Marco N, Clemente-Ciscar M, Cordero T, Muñoz-Blat I, Amadoz A, et al. Disrupted pgr-b and esr1 signaling underlies defective decidualization linked to severe preeclampsia. *Elife*. 2021;10. doi:10.7554/eLife.70753
209. Li T, Li X, Guo Y, Zheng G, Yu T, Zeng W, et al. Distinct mRNA and long non-coding RNA expression profiles of decidual natural killer cells in patients with early missed abortion. *FASEB J*. 2020;34: 14264–14286. doi:10.1096/fj.202000621R
210. Chen P, Zhou L, Chen J, Lu Y, Cao C, Lv S, et al. The Immune Atlas of Human Deciduas With Unexplained Recurrent Pregnancy Loss. *Front Immunol*. 2021;12: 689019. doi:10.3389/fimmu.2021.689019
211. Pierre-Jean M, Deleuze J-F, Le Floch E, Mauger F. Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. *Brief Bioinform*. 2019. doi:10.1093/bib/bbz138
212. Qi R, Ma A, Ma Q, Zou Q. Clustering and classification methods for single-cell RNA-sequencing data. *Brief Bioinform*. 2019;21: 1196–1208. doi:10.1093/bib/bbz062
213. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimed Tools Appl*. 2019. doi:10.1007/s11042-018-6894-4
214. Davies DL, Bouldin DW. A Cluster Separation Measure. *IEEE Trans Pattern Anal Mach Intell*. 1979;PAMI-1: 224–227. doi:10.1109/TPAMI.1979.4766909
215. Caliński T, Harabasz J. A Dendrite Method For Cluster Analysis. *Commun Stat*. 1974. doi:10.1080/03610927408827101
216. Hassani M, Seidl T. Using internal evaluation measures to validate the quality of diverse stream clustering algorithms. *Vietnam J Comput Sci*. 2017;4: 171–183. doi:10.1007/s40595-016-0086-9
217. Hubert L, Arabie P. Comparing partitions. *J Classif*. 1985;2: 193–218. doi:10.1007/BF01908075
218. Amigó E, Gonzalo J, Artilles J, Verdejo F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf Retr Boston*. 2009;12: 461–486. doi:10.1007/s10791-008-9066-8
219. Rand WM. Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc*. 1971. doi:10.1080/01621459.1971.10482356
220. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. *Nat Biotechnol*. 2019;37: 547–554. doi:10.1038/s41587-019-0071-9
221. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One*. 2010;5: e12776. doi:10.1371/journal.pone.0012776
222. Rabaglino MB, Uiterweer EDP, Jeyabalan A, Hogge WA, Conrad KP. Bioinformatics approach reveals evidence for impaired endometrial maturation before and during early pregnancy in women who developed preeclampsia. *Hypertension*. 2015;65: 421–429. doi:10.1161/HYPERTENSIONAHA.114.04481
223. Garrido-Gomez T, Dominguez F, Quiñonero A, Diaz-Gimeno P, Kapidzic M, Gormley M, et al. Defective decidualization during and after severe preeclampsia reveals a possible maternal contribution to the etiology. *Proc Natl Acad Sci U S A*. 2017;114: E8468–E8477. doi:10.1073/pnas.1706546114
224. Queckbörner S, von Grothusen C, Boggavarapu NR, Francis RM, Davies LC, Gemzell-Danielsson K. Stromal heterogeneity in the human proliferative endometrium—a single-cell rna sequencing study. *J Pers Med*. 2021;11. doi:10.3390/jpm11060448
225. Rawlings TM, Makwana K, Taylor DM, Molè MA, Fishwick KJ, Tryfonos M, et al. Modelling the impact of decidual senescence on embryo implantation in human endometrial assembloids. Spencer TE, Cooper JA, Spencer TE, Wagner G, Vankelecom H, Julie Kim J-Y, editors. *Elife*. 2021;10: e69603. doi:10.7554/eLife.69603

226. Menéndez ML, Pardo JA, Pardo L, Pardo MC. The Jensen-Shannon divergence. *J Franklin Inst.* 1997;334: 307–318. doi:10.1016/s0016-0032(96)00063-4
227. Pavličev M, Wagner GP, Chavan AR, Owens K, Maziarz J, Dunn-Fletcher C, et al. Single-cell transcriptomics of the human placenta: Inferring the cell communication network of the maternal-fetal interface. *Genome Res.* 2017;27: 349–361. doi:10.1101/gr.207597.116
228. Grasso E, Gori S, Soczewski E, Fernández L, Gallino L, Vota D, et al. Impact of the Reticular Stress and Unfolded Protein Response on the inflammatory response in endometrial stromal cells. *Sci Rep.* 2018;8: 12274. doi:10.1038/s41598-018-29779-8
229. Masat E, Gasparini C, Agostinis C, Bossi F, Radillo O, De Seta F, et al. RelB activation in anti-inflammatory decidual endothelial cells: a master plan to avoid pregnancy failure? *Sci Rep.* 2015;5: 14847. doi:10.1038/srep14847
230. Wang D, Malarkannan S. Transcriptional regulation of natural killer cell development and functions. *Cancers (Basel).* 2020;12: 1–33. doi:10.3390/cancers12061591
231. Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics.* 2018. doi:10.1093/biostatistics/kxx017
232. Chalise P, Fridley BL. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PLoS One.* 2017. doi:10.1371/journal.pone.0176278
233. Nguyen H, Shrestha S, Draghici S, Nguyen T. PINSPPlus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics.* 2019. doi:10.1093/bioinformatics/bty1049
234. Ramazzotti D, Lal A, Wang B, Batzoglou S, Sidow A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nat Commun.* 2018. doi:10.1038/s41467-018-06921-8
235. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. *Nature.* 2012;490: 61–70. doi:10.1038/nature11412
236. Cabassi A, Kirk PDW. Multiple kernel learning for integrative consensus clustering of omic datasets. *Bioinformatics (Oxford, England).* 2020. pp. 4789–4796. doi:10.1093/bioinformatics/btaa593
237. Nguyen T, Tagett R, Diaz D, Draghici S. A novel approach for data integration and disease subtyping. *Genome Res.* 2017;27: 2025–2039. doi:10.1101/gr.215129.116
238. Huh R, Yang Y, Jiang Y, Shen Y, Li Y. SAME-clustering: Single-cell Aggregated Clustering via Mixture Model Ensemble. *Nucleic Acids Res.* 2020;48: 86–95. doi:10.1093/nar/gkz959
239. Singh R, Narayan A, Hie B, Berger B. Schema: A general framework for integrating heterogeneous single-cell modalities. *bioRxiv.* 2019. doi:10.1101/834549
240. John CR, Watson D, Barnes MR, Pitzalis C, Lewis MJ. Spectrum: fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics.* 2019;36: 1159–1166. doi:10.1093/bioinformatics/btz704
241. Welch JD, Hartemink AJ, Prins JF. MATCHER: Manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol.* 2017;18. doi:10.1186/s13059-017-1269-0
242. Amodio M, Krishnaswamy S. MAGAN: Aligning biological manifolds. 35th International Conference on Machine Learning, ICML 2018. 2018. pp. 327–335.
243. Cao K, Bai X, Hong Y, Wan L. Unsupervised Topological Alignment for Single-Cell Multi-Omics Integration. *Bioinformatics.* 2020. doi:10.1101/2020.02.02.931394
244. Liu J, Huang Y, Singh R, Vert JP, Noble WS. Jointly embedding multiple single-cell omics measurements. *Leibniz International Proceedings in Informatics, LIPIcs.* 2019. doi:10.4230/LIPIcs.WABI.2019.10
245. Lock EF, Dunson DB. Bayesian consensus clustering. *Bioinformatics.* 2013;29: 2610–2616. doi:10.1093/bioinformatics/btt425

246. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*. 2012;28: 3290–3297. doi:10.1093/bioinformatics/bts595
247. Wang X, Sun Z, Zhang Y, Xu Z, Xin H, Huang H, et al. BREM-SC: a bayesian random effects mixture model for joint clustering single cell multi-omics data. *Nucleic Acids Res*. 2020;48: 5814–5824. doi:10.1093/nar/gkaa314
248. Campbell KR, Steif A, Laks E, Zahn H, Lai D, McPherson A, et al. Clonealign: Statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome Biol*. 2019;20. doi:10.1186/s13059-019-1645-z
249. Martin C, Welch J, Kozareva V, Ferreira A, Vanderburg C, Martin C, et al. Integrative inference of brain cell similarities and differences from single-cell genomics. *bioRxiv*. 2018. doi:10.1101/459891
250. Duren Z, Chen X, Zamanighomi M, Zeng W, Satpathy AT, Chang HY, et al. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc Natl Acad Sci U S A*. 2018;115: 7723–7728. doi:10.1073/pnas.1805681115
251. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol*. 2020;21. doi:10.1186/s13059-020-02015-1
252. Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, et al. Joint probabilistic modeling of paired transcriptome and proteome measurements in single cells. *bioRxiv*. 2020. doi:10.1101/2020.05.08.083337
253. Zuo C, Chen L. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Brief Bioinform*. 2020. doi:10.1093/bib/bbaa287
254. Yang KD, Belyaeva A, Venkatachalapathy S, Damodaran K, Katcoff A, Radhakrishnan A, et al. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nat Commun*. 2021;12. doi:10.1038/s41467-020-20249-2
255. Gabasova E, Reid J, Wernisch L. Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. *PLoS Comput Biol*. 2017;13. doi:10.1371/journal.pcbi.1005781
256. Fernandes N, Gkolia A, Pizzo N, Davenport J, Nair A. Unification of HDP and LDA Models for Optimal Topic Clustering of Subject Specific Question Banks. 2020 [cited 28 Mar 2023]. Available: <https://arxiv.org/abs/2011.01035v1>
257. Muter J, Kong C-S, Brosens JJ. The Role of Decidual Subpopulations in Implantation, Menstruation and Miscarriage. *Front Reprod Heal*. 2021;3: 804921. doi:10.3389/frph.2021.804921
258. Leitao B, Jones MC, Fusi L, Higham J, Lee Y, Takano M, et al. Silencing of the JNK pathway maintains progesterone receptor activity in decidualizing human endometrial stromal cells exposed to oxidative stress signals. *FASEB J*. 2010;24: 1541–1551. doi:10.1096/fj.09-149153
259. Kajihara T, Jones M, Fusi L, Takano M, Feroze-Zaidi F, Pirianov G, et al. Differential expression of FOXO1 and FOXO3a confers resistance to oxidative cell death upon endometrial decidualization. *Mol Endocrinol*. 2006;20: 2444–2455. doi:10.1210/me.2006-0118
260. Tang JJJ, Sung AP, Guglielmo MJ, Navarrete-Galvan L, Redelman D, Smith-Gagen J, et al. Natural killer (Nk) cell expression of cd2 as a predictor of serial antibody-dependent cell-mediated cytotoxicity (adcc). *Antibodies*. 2020;9: 1–18. doi:10.3390/antib9040054
261. Dimitriu MA, Lazar-Contes I, Roszkowski M, Mansuy IM. Single-Cell Multiomics Techniques: From Conception to Applications. *Front Cell Dev Biol*. 2022;10. doi:10.3389/fcell.2022.854317
262. Rabaglino MB, Conrad KP. Evidence for shared molecular pathways of dysregulated decidualization in preeclampsia and endometrial disorders revealed by microarray data integration. *FASEB J*. 2019;33: 11682–11695. doi:10.1096/fj.201900662R



**TURUN
YLIOPISTO**
UNIVERSITY
OF TURKU

ISBN 978-951-29-9449-6 (PRINT)
ISBN 978-951-29-9450-2 (PDF)
ISSN 2736-9390 (Painettu/Print)
ISSN 2736-9684 (Sähköinen/Online)