CLASSIFICATION OF PERIPHERAL BLOOD MONONUCLEAR CELLS USING SINGLE CELL TRANSCRIPTOMICS DATA AND ARTIFICIAL NEURAL NETWORKS

JIAHUI ZHONG

Thesis submitted to The University of Nottingham for the degree of Doctor of Philosophy

NOVEMBER 2022

‘We must ascribe to all cells an independent vitality.’

Theodor Schwann, 1810 - 1882.

ACKNOWLEDGEMENTS

This thesis means the full stop of my PhD journey - the five years of my entire life, that has taught me about changing, becoming, and, being.

First of all, I must express my deepest gratitude to my principal supervisor, Prof. Vladimir Brusic, for your invaluable guidance, generous patience, unwavering support, unusual wisdom, encouraging optimism, and your precious time, throughout my whole PhD study journey.

I will be forever grateful for your resolute belief in me and constant compassionate care for me. There have been many difficult times, but your guidance has always been leading me and encouraging me as the lighthouse on the sea. The road to the mountain of books, is named the diligence, the boat in the sea of knowledge, is named the bitterness. Thank you for always sailing me back to the course and guiding me to reach the destination, when I was indulged in the colorful novelties of science. Thank you for tolerating my endless divergent thinking and consistently leading me to focus on the scope of my study. Thank you for sharing me the opportunity getting involved in this beautiful idea, the one refers to the start of predictive health in single cell era, solely scientists who have broad vision and great love for human beings could have drawn the picture of it - I often wondered how the idea was conceived. Thank you for accommodating my anxieties and emotions in those difficult times and imbuing me with your optimism and fortitude to keep moving forward. Thank you for sharing me all your knowledge and experience without reservation. Thank you for explaining those incomprehensible concepts thousands of times for me. Thank you for being so accessible, humorous, and frank when sharing your considered advice, thank you for your generous heart, wisdom, and perspective vision to confront the many challenges I brought your way. Thank you for your enormous precious time spending on guiding me and helping me to be more mature, both in life and in scientific research.

My heartfelt gratitude to my beloved research project, for your unbelievable attraction to me. You showed me how human life could be like after many years later. I felt very grateful having you as my PhD topic. I hope there would be more people knowing the power of you and making human society better.

My sincere thanks extend to all my colleagues, Razin Shaikh, Haoguo Wu, Minjie Lyu, Sen Lin, Xin Lin, Yihan Zhang, Luning Yang, and everyone, for your generous help that have been essential for the work.

I would like to thank Prof. Huan Jin and Prof. Heshan Du for being my co-supervisors and sharing valuable advice and generous support all the time.

I would like to thank Prof. Anthony Bellotti and Prof. Saeid Ardakani as my internal assessors, for your valuable and constructive comments on this work, that make us see the project from different perspectives.

I would like to thank all our collaborators, Prof. Guanglan Zhang, Prof. Derin Keskin, Prof. Nenad Mitic, and Prof. Zhiwei Cao, for your generous help and support to the work.

I would like to thank every academic and administrative staff of Graduate School and Faculty of Science and Engineering, for your kind support in this process.

Further thanks I would like to extend to all my sweet and optimistic friends, Zhao Liu and Shuai Cheng, for your generous encouragement and support throughout the journey.

My gratitude to all the people who I have met along the way. You have made me understand the world and myself more truly and have made me understand the meaning of human pursuit of science and truth.

Importantly, I would like to thank my friend, Gang Xiao, for your great support and encouragement during my thesis writing period.

Finally, and the most, this thesis is dedicated to my dearest, darling mother and father, for all your pure love, boundless support, and full encouragement unconditionally in everything I do.

Thanks to all the souls, the time, and the destiny.

ABSTRACT

This thesis presents our research on single cell classification with single cell transcriptomics (SCT) data and purely supervised machine learning (ML) method artificial neural network (ANN).

SCT sequencing technology can accurately capture the instantaneous gene expression of every single cell. The 10x SCT technology has realized SCT profiling in a high-throughput and cost-efficient manner. It can produce over 10^9 transcripts of over 10^5 individual cells with ~33,000 gene features, for profiling a targeted sample in a single study. However, the classification of single cells with SCT data has met challenges. These include: the lack of supervised ML methods in single cell classification, the lack of reference datasets for SCT gene expression profiles, the lack of a specific cell ontology for single cell classification, the characteristic of SCT data - large data size, high-dimensional, the sparsity (a large proportion of zero-counts), and the presence of variables (biological and technical). The currently used unsupervised ML methods have shown the limitation on generalization and manual inspection to annotation.

In addressing the needs and challenges, considering the capability of generalization and the suitability to large data size, high-dimensional, sparse, and high-variety SCT data, we made the hypothesis that single cell classification can be done with the supervised ML method ANN and SCT data. We selected peripheral blood mononuclear cells (PBMC) as the SCT data sample for this study. PBMC is a conventionally used predictive health indicator, it has five main cell types that are naturally isolated. The accurate classification of SCT data of the five cell types can be used in early disease diagnosis and the realization of accurate blood testing based on SCT analysis.

We prepared standardized 56 reference datasets for PBMC SCT classification and described a multi-dimensional cell ontology with over 163 dimensions for single cell classification, with PBMC as an example.

In the initial study, the proof of concept that using the supervised ML method ANN and standardized SCT data to realize single cell classification has been demonstrated, with an overall accuracy of 89.4%. Follow-up, we deployed holdout internal cross-validation, external validation, added data validation, together with cyclical incremental learning method, and newly collected independent SCT datasets from four sources, to investigate the baseline for highly accurate PBMC SCT classification. The overall accuracy of the 4-class classification was 93.0%, and the 5-class classification achieved 94.6%. The classification results have been analyzed with PBMC SCT cell ontology and basic statistics. B cells, monocytes, and T cells had classification accuracy that was greater than 95%. Due to similarities between NK cells and T cell subsets, the classification accuracy of NK cells was maintained at roughly 75%. The accuracy of dendritic cells was limited

due to the small proportion of numbers in the training sets.

Based on these, we studied the effect of various processing protocols of SCT data on single cell classification. The findings indicated that datasets from samples with minimally processing protocols (PBMC separation only) helped in the identification of SCT gene expression patterns.

Further, we explored the vulnerability of ANN-SCT-PBMC classifiers, using 17 non-representative datasets of five different confounding factor groups, and 17 rounds of cyclical four-supersets-swapping external validation experiments. The results revealed that when trained with sufficient reference datasets, the ANN-SCT-PBMC model was robust and could survive a small number of non-representative instances hidden in the training set. The model can recognize and assess the representativeness of SCT data once it has been trained on purified high-quality reference data. The proportions of reference and non-representative datasets, the distribution of classes in training and testing sets, the similarity of gene expression between cell types and subtypes, the characteristics of non-representative datasets, etc. are variables that had an impact on model vulnerability.

This research gives a solution to the current "eleven grand challenges" of SCT data analysis. It demonstrates that purely supervised ML ANN is a viable option for classifying cell types from single cell expression data, with generalization capability and robustness on various upcoming data sets. This research reveals that sufficient reference SCT data, generated with precise and strict protocols and labeled with a complete and detailed multi-dimensional cell ontology, is required for highly accurate single cell classification, that can contribute to future predictive health development and hematology development.

KEY WORDS: single cell classification, single cell transcriptomics (SCT) data, supervised machine learning (ML), artificial neural network (ANN), peripheral blood mononuclear cells (PBMC), multi-dimensional cell ontology, proof of concept, incremental learning, model vulnerability, data representativeness, model robustness.

LIST OF ABBREVIATIONS

10x	10x Genomics Demonstration
ANN	Artificial Neural Network
ACC	Accuracy
BC	B Cells
CL	Cell Ontology
DC	Dendritic Cells
F1	F1-Score
FACS	Fluorescence-Activated Cell Sorting
FN	False Negative
FP	False Positive
GEO	Gene Expression Omnibus
iNKT	iNKT (invariant Natural Killer T Cells)
MACS	Magnetic-Activated Cell Sorting
MAIT	Mucosal-Associated Invariant T Cells
MC	Monocytes
ML	Machine Learning

NK	Natural Killer Cells
NKT	Natural Killer T Cells
PBMC	Peripheral Blood Mononuclear Cells
pDC	plasmacytoid Dendritic Cells
SCT	Single Cell Transcriptomics
SE	Sensitivity
SOP	Standard Operating Procedures
SP	Specificity
TC	T Cells
TN	True Negative
TP	True Positive
Vd1	Gamma-delta (γδ) 1 T Cells
Vd2	Gamma-delta (γδ) 2 T Cells

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

1.1 Background

The bulk transcriptomics sequencing technology measures average gene expression value of mixed biological samples. The unique heterogeneity of individual single cell cannot be characterized, that leads to the loss of important genetic information. Currently, single cell transcriptomics (SCT) sequencing technology has been developed, it can capture and reveal unique gene expression of individual single cell, that detects cell heterogeneity and refines existing cell ontology. It can be used in predictive health and early disease diagnosis. The 10x Genomics high-throughput SCT sequencing platform has clear and standardized experimental procedures that produce reliable and consistent SCT data in batches.

SCT data has great value to human health and life science. However, the high-dimensionality, high sparseness, dropouts, biological variables and technical variables of SCT data make the classification of SCT data a challenge [1]. Currently, unsupervised machine learning methods such as principal component analysis (PCA) and clustering have been used to classify cells with SCT data, but it has demonstrated weak robustness, accuracy, and sensitivity when it comes to multi-source data from different independent studies [2]. The value of SCT data cannot be used fully by unsupervised machine learning methods, that cannot generalize on various SCT data sets of different independent sources. We consider to use supervised machine learning method artificial neural network (ANN) to solve the challenges of single cell classification with high dimensional SCT data.

Peripheral blood mononuclear cells (PBMC) is the significant research objective for human health status detection, disease diagnosis, the development of immunology research, cancer research and toxicology applications. The cell type, cell status and cell number of PBMC in an individual body indicate the selective responses of immune system.

This study has made the hypothesis and tried to prove the concept that single cell classification can be done with SCT data and supervised machine learning method ANN, with satisfied and practically applicable accuracy. To build, prove, and study the prototype of supervised ANN classification model in PBMC SCT pattern recognition, can make efficient use of exponentially growing SCT data, and demonstrate the concept of data-based predictive health with PBMC SCT gene expression profiles.

1.2 Motivation & Hypothesis

1.2.1 The importance of SCT technology

Single cell transcriptomics sequencing (SCT or scRNA-seq) technology detects gene expression profiles of individual single cells in a biological sample. Gene expression by bulk sequencing from mixed samples provides only average gene expression across all cells in the sample. SCT preserves information about the heterogeneity of gene expression within cell types and subtypes and their various states [3]. Data sets from SCT studies are in form of sparse matrices having >30,000 genes (features) in rows, and up to 100,000 cells in matrix columns. These data sets are growing at an exponential rate both in the number of cells per matrix, and in the number of data sets that are available for analysis [4, 5].

Classification of single cells is essential for analyzing the composition of tissues and the cellular basis of health and disease status. Accurate classification of cell types and subtypes, along with the identification of their gene and protein expression patterns, enable understanding to cellular and molecular basis of biological processes [6]. The differences between healthy and disease states are reflected in differential gene expression, it allows for medical applications of single cell technologies: diagnostic and prognostic applications, and disease treatment selection [7] in cancer, infectious disease, autoimmunity, and other pathological states [8].

The first report of single cell gene expression was published in 2009 [9]. Major breakthroughs in microfluidics and cell labeling methods have enabled high-throughput of single cells, rapid standardized SCT gene expression measurement, and analysis [4, 5, 10]. The conventional classification rules applied to cell populations are mainly qualitative and are based on lineage, phenotypic markers, and simple, functional properties [11]. The SCT uses gene expression and quantitative methods to define cell types and precisely describe their lineage, phenotype, function, and various states [11]. Such cellular gene expression profiles and their variants (due to different sample processing methods) are cataloged in single cell atlases [12, 13]. Bulk-sequencing methods produce mean gene expression values of millions of cells. In contrast, SCT produces gene expression profiles characteristic of cell sets defined by a much finer grouping of cells that share origin, function, subtype, and biological status [3].

1.2.2 The 10x Genomics platform

The 10x Genomics SCT sequencing platform scaled up to enable routine measurements of expression count over 10^5 cells with ~33,000 gene features in a single study that produces over 10^9 transcript counts values profiling a targeted sample [10, 14].

It combines high throughput (up to 40,000 cells in a single experiment), high cost-efficiency, and rapid turnaround (1-2 days from sample collection to results) [15]. When the cell viability is greater than 90%, the cell capture rate of one single sample can reach 65% (10x protocol). The 10x SCT data is represented by a high-dimensional sparse matrix. A single cleaned 10x SCT data set (sparse matrix) can have 10^9-10^{10} data points because it has up to 10^5 columns representing individual cells and >30,000 rows representing features (gene counts). It has observed that 90-99% of the values are zero [16].

The 10x SCT has formed strict standard experimental procedures that can produce highly reproducible measurements, even in samples from different individuals. The available capture probes provide high coverage of the genome. 10x was benchmarked against several alternative methods [17, 18] and it is emerging as a popular SCT platform.

High throughput SCT is a prototypic big data technology. Since 2017, with the emergence of the 10x Genomics platform, the large-scale unified 10x scRNA-seq data sets have been generated and have grown exponentially with more than 52,500 10x data sets available in GEO data repository [19] (www.ncbi.nlm.nih.gov/geo), as of May 2023.

Currently, the analysis of 10x SCT data focuses on single cell annotation and classification aimed at understanding biological mechanisms, such as cellular differentiation, tissue distribution of cells, the discovery of new biomarkers, detection of rare cell types, assessment of tumor heterogeneity, detecting gene activation pathways related to pathology, and detecting molecular and cellular responses to therapeutic interventions [20-22].

1.2.3 The challenges and difficulties in SCT data analysis

The single cell classification and adequate utilization of SCT data has been a **challenge** to researchers for a long time [1]. **First, high sparsity.** 10x SCT generates large but sparse matrices (over 95-99% of values are typically zeros, that depends on the depth of sequencing implemented and the internal expression level of gene features. It can perplex and obstacle the following

downstream analysis. The zero value is attributed to true zero value (the gene is not expressed in the cell at this transient moment) or "dropout" phenomenon (the transcript is not captured). **Second, high variety.** In 10x SCT profiles, there can be errors and noises, such as multiplets (doublets or triplets, when two or three single cells are wrapped in one oil droplets), and bias values resulting from biological (sample conditions – fresh/ frozen thawed, activated status, stimulated status) or technical (chemical reagent, machine version, batch effect, etc.) confounding factors.

Third, high dimensionality. There are >30,000 dimensions in the gene list. Efficiently preserving valuable information during analyzing high dimensional (>30,000 features) SCT expression data matrix with $>10^5$ cell numbers has not formed an acknowledged approach so far. **Forth, multiple sources and integration.** The current-in-use SCT data analysis pipelines meet difficulties to integrate and generalize the stylized analysis protocols to SCT data that has been sequenced with multi sample preparing procedures and diverse experimental measurement conditions. It is difficult to analyze SCT data collected from various sources (different studies and labs). It involves batch effect and various features in gene list (features are various in data set of different study and different source). **Fifth, lack of reference data sets and single cell ontology.** The classification of single cells lacks precise expression profile definitions and sufficient reliable standard references [1]. There is currently no available standardized reference dataset for single cell classification. Also, there is an urgent requirement for a single cell ontology as reference to categorize single cells from multiple dimensions [23].

1.2.4 The importance of PBMC classification

Peripheral blood mononuclear cells (PBMC) are circulating immune cells with a single round nucleus in the blood and are common diagnostic and prognostic targets [24]. PBMC are composed of mixed cell populations. There are five main subtypes of PBMC: B cells (BC), monocytes (MC), dendritic cells (DC), T cells (TC) and natural killer cells (NK) [25]. Frequencies of PBMC subtypes can vary widely from individual to individual, but also over time within the same individual [26]. A rough consensus over multiple antibody catalogue estimates is that B cells make 5-15%, monocytes make 10-30%, DC make 1-2%, NK cells make 5-10%, and T cells make 40-70% of PBMC in humans [25]. Normal ranges (reference values) of the numbers of specific cell types or subtypes in PBMC vary by 5 to 20 folds in healthy individuals [27]. Their transcriptome profiles show high variation, primarily resulting from sample processing steps [28] and the health/disease status of the tissue [24, 26, 29]. Gene expression profiles in PBMC that circulate in blood were shown to be different from the tissue resident PBMC [16]. This suggests that gene

expression differences can also be used to identify the tissue of origin of resident PBMC [30].

PBMC has been extensively used in the study of infectious disease, immunology and autoimmunity, transplantation, oncology, and vaccine development. PBMC are important targets of single-cell studies because they are indicators of immune status and are studied in cell function, transcriptional regulation, identification of biomarkers, and disease modeling [31-33], pharmacogenomics [31, 34], hematological malignancies, among others [35-37]. PBMC are routinely used for monitoring health and for the diagnosis of infection and blood disease [38-40].

PBMC cell type has characteristic patterns of gene expression that is determined by multiple factors. These factors include the cell differentiation stage, tissue and organ localization, developmental stage, epigenetic modification, activation status, age, health/disease status, and other factors [26, 41]. Final differentiated cell types emerge through molecular changes of developmental pathways characterized by recognizable patterns of gene-expression and protein markers [42].

There is a need in a single cell ontology for PBMC classification. Hundreds of subtypes have been described in literature, but unified ontology of PBMC does not exist [43]. Subsets of PBMC are identified through analysis of their surface receptors by flow cytometry [44] or by analyzing their transcriptomics profiles [45]. More than 120 cell subsets of PBMC have been described [46], but current descriptions of PBMC subsets are incomplete and the efforts to define them are ongoing [47, 48].

In addition to the inherent biological differences, each step in the process of peripheral blood sampling, storage, preparation, and measurement as well as their duration will change gene expression in single cells [49-51]. At present, uniform and strict standards have been established for sample collection, preparation, and storage of PBMC [49, 52], to ensure yield, viability and preservation of function [53, 54]. Also, PBMC is naturally isolated, that minimizes external stimuli during tissue isolation and cell sorting procedure. These largely preserves specific gene expression profiles of PBMC under individual circumstance [53]. Standard operating procedures (SOP) have been defined and established for the latest single cell transcriptomics (SCT) technologies [55], enabling the improved reproducibility of SCT studies. The combination of advanced SCT technologies and the rapidly increasing availability of data sets provide a basis for defining cell types and subtypes by SCT gene expression profiles from diverse datasets.

Specific PBMC profile done with 10x SCT sequencing can represent the differences in gene expression of immune cells referring to each individual body [38]. Regular monitoring and comparative analysis of PBMC components and the frequency of each component can realize the understanding of human health and disease prevention and diagnosis [39, 40]. The cell

classification and cell counting of PBMC sample can be completed by fluorescence-activated cell sorting (FACS). However, the realization of low cost and high efficiency blood monitoring and analysis requires the establishment of a computerized PBMC sample cell classification system through single cell sequencing technology and machine learning technology.

1.2.5 The limitation of unsupervised ML methods

The characteristics of SCT data – large size, sparseness, sensitivity to sample processing and experimental conditions, biases and random errors in data, and lack of reference data sets – require advanced statistical and **machine learning (ML)** techniques essential for the analysis of sparse matrices (downstream analysis).

SCT data sets are produced using various sample processing conditions and they represent many different biological states, making SCT data highly heterogeneous. The lack of reference data sets mandates the use of **unsupervised ML approaches** [22], predominantly unsupervised clustering [22]. Unsupervised ML methods are broadly used for labeling and classification of single cells either alone [56] or in combination with supervised ML methods [57]. Unsupervised ML methods deploy a combination of clustering algorithms to group single cells together, with semi-automated labeling and manual annotation [22, 58] based on marker genes.

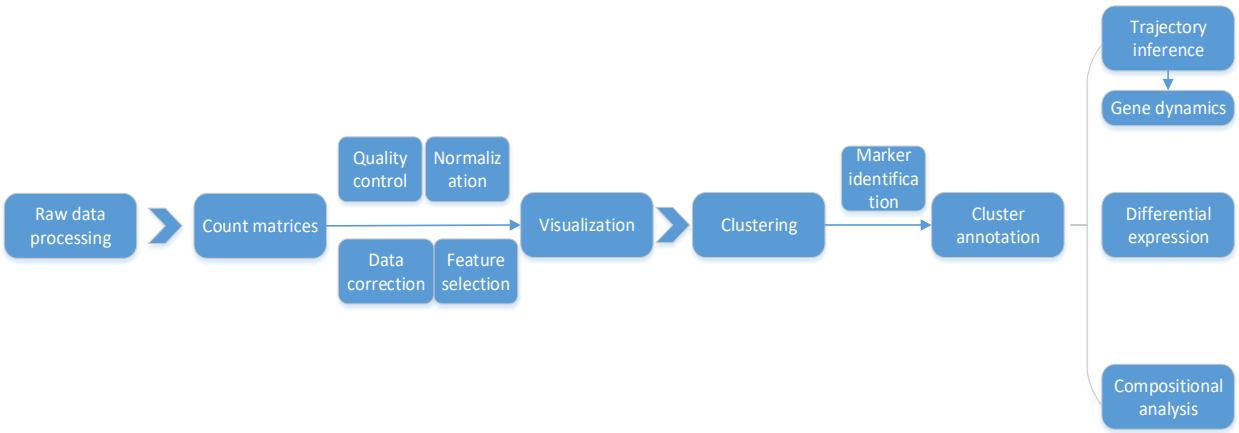


Figure 1. A typical SCT analysis workflow using unsupervised ML for one study at a time. After data pre-processing, single cells with SCT profiles are grouped with unsupervised clustering methods and annotated with significant marker genes, manually and empirically.

However, the number of classes in unsupervised methods is unknown – it is estimated by identified clusters and biological interpretation [2]. Also, the marker genes used are manually defined. These both introduce subjective judgments and different expert opinions (knowledge bias). Further, unsupervised ML methods do not scale up well, and the workflows lack generalization – solely typically applied to some specific dataset of mixed-class cells – a workflow that performs well on a specific dataset does not perform well on datasets produced from different studies [22, 56, 57] (insufficient robustness, reproducibility and sensitivity for multi-source data sets).

Several bottlenecks currently limit the analysis to the tools of unsupervised ML, including the lack of standardized formats for data sets, lack of reference gene expression profiles, high-dimensional nature of data, the sparsity of data (large proportion of zero-counts), and presence of noise in data (errors and biases). On the other hand, the SCT gene expression of the same sample, when sample processing procedure and experimental conditions are standardized, are highly reproducible [18, 59]. A semi-supervised method that used variational autoencoder neural network architecture was reported to outperform unsupervised methods, that demonstrates the trend of applying supervised learning method for cell classification of SCT data [60].

1.2.6 The hypothesis of using supervised ML method ANN

Supervised ML method can support as a solution to solve the challenges of studying and analyzing SCT data. It is expected to have superior generalization ability and performance on single cell classification across different studies, making accumulated SCT data comparable and valuable. Supervised ML classification systems use algorithms that are logic-based (such as decision trees, rule-based classifiers), network-based (such as artificial neural networks, support vector machines), statistic-based (Bayesian algorithms), or instance-based (such as distance-based or pattern recognition methods) [6]. Supervised ML can perform classification using single-cell gene expression profiles across various studies representing diverse sample processing conditions and experimental settings.

Supervised learning method **artificial neural networks (ANN)** [61] can be used for advanced SCT cell classification. Compared to other supervised ML methods, ANN is efficiently suitable for task with a large scale of complex training data [62].

ANN fits to deal with the complexity of SCT data: large data size (>10,000 observations in one dataset); high-dimensional features (>30,000); full of variables (biological/technical); sparse matrix (>90% zeros); multiple sources (data collected from different studies).

It is convenient to implement, especially with high-dimensional noisy data that has unknown mathematical relationships in features. It has the capability in capturing nonlinear and complex underlying characteristics in SCT profiles, with high degree of accuracy [63].

ANN can address complexity, and it is regularly among the most performing [63]. ANN allows to solve the problem with incomplete knowledge [64], it can be used as the first approach to prototype. It is data-driven, adaptive learning and self-organization, that learns tasks based on given data for training and creates its own representation of the information [63].

ANN can learn the full features of each instance and make prediction decision. In SCT data, each feature can be important to single cell pattern recognition, ANN can ensure the integrity of training information and ensure full-dimensional learning (rather than dimensionality reduction). It learns to recognize the full internal patterns that exist in the data [63].

Further, ANN can be sensitive and flexible to changing environment [63] (e.g. tiny gene expression pattern changes in over 30,000 features [65]). ANN is adaptive to constantly changing input for complex and exponentially growing SCT data – where the relationships are quite dynamic and non-linear. It is convenient to observe the behavior of model on data effect. This project tries to study and understand the influence of SCT data to model behavior. The factors include data sources, data generation conditions, and other dimensions in a multi-dimensional cell ontology.

Thus, from the above aspects, we have the motivation and hypothesize to use ANN for the SCT classification task. In principle, all tasks can be solved with various supervised ML methods, including support vector machine (SVM), random forest (RF), etc. While SVM is suitable for tasks with a small amount of training information and regular binary classification, and RF can take risks in overfitting. For the SCT classification task – with large-scale, high dimensional, high sparsity, complex, and variable data, and it requires satisfied robustness, we consider ANN is the first choice to perform the prototype verification.

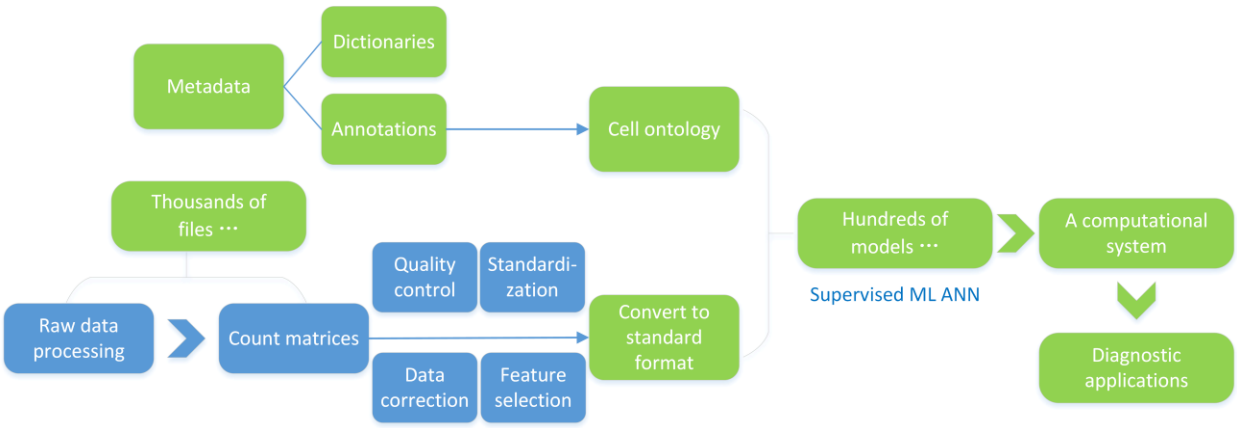


Figure 2. This project's single-cell RNA-seq analysis workflow using supervised ML method ANN.

Computerized SCT cell classification using ANN can bring purely supervised, specific labeled learning and classification procedure to each individual single cell gene count expression profile, where is improved to unsupervised ML clustering and biological manual cell sorting FACS. ANN algorithms extract original features from large annotated SCT data sets and use them to create a prediction tool based on hidden patterns. Once the training is completed, the algorithm can apply this training to analyze other data, that generalizes the learning and classification procedure to multi source data sets with diverse experimental conditions. Exact, specific, clean annotation of SCT data sets is required for ANN model training and cell type prediction.

Currently, there is no purely supervised ML method implemented, because there is no reference data available. The main aim of the project is to demonstrate and prove the concept that single cell classification can be done with SCT data and supervised ML method ANN. It aims to build and demonstrate a prototype and a protocol to use supervised ML to handle high-dimensional, noisy, large size SCT data, solving the difficulties in Eleven Grand Challenges [22] – correctly classifying and labeling single cells in SCT data with prepared reference data sets. PBMC classification with SCT data and ANN aims to build purely supervised classification prototype of SCT, observe data effects from multi-dimensional PBMC-SCT cell ontology, and be potentially useful in early diseases diagnosis and predictive health.

1.3 Goal & Objectives

1.3.1 Overall goal

The overall goal is to prove a concept that we can do highly accurate classification of blood cells using SCT data and supervised ML method ANN, this method must be highly accurate, must generalize well across different studies, it must be applicable in practice and in real life. The analysis of SCT data with supervised ML method can help to solve several questions in the "eleven grant challenges" of SCT data analysis that have been listed in an article of 2020 [22]. The classification model should take good use of SCT data and reveal the specific gene expression profile of individual cell type, with observation of data quality and data effects (multiple dimensions in PBMC-SCT cell ontology) to ANN model behavior.

1.3.2 Specific objectives

1. **Organize the data**
    a. Select relevant data sets, convert them into standardized format ready to analyze, and perform quality control.
    b. Update the common list of genes ("gene common list") for comparative analysis. Gene common list should be prepared for standardization conversion process.
    c. Establish experimental and statistical metadata for data sets that have study description information and summary basic statistical information.
    d. Cell ontology preparation for involved 10x SCT data sets.

2. **Prove the concept**
    a. Prove the concept that computerized simulation of PBMC classification can be accomplished with SCT data and purely supervised ML method ANN.

3. **Data accumulation incremental learning**
    a. Prepare a certain amount of clean and standardized SCT data sets to train ANN model using incremental learning method (data accumulation), trying to study the accuracy, sensitivity, and specificity of ANN classification model simulating real life situation.
    b. ANN should perform robustness across different data sets with different sources, different experimental platforms, and different experimental conditions.

4. **Data representativeness and data effect**
    a. Observe the data representativeness and data effects (multiple dimensions in PBMC-SCT cell ontology) to ANN model behavior, analyzing the importance of data effects to single cell classification.

1.4 Overall Study Design

The overall design of this project is to use PBMC SCT datasets generated from 10x technology and supervised ML method ANN to demonstrate purely supervised SCT single cell classification (as shown in

Figure 3).

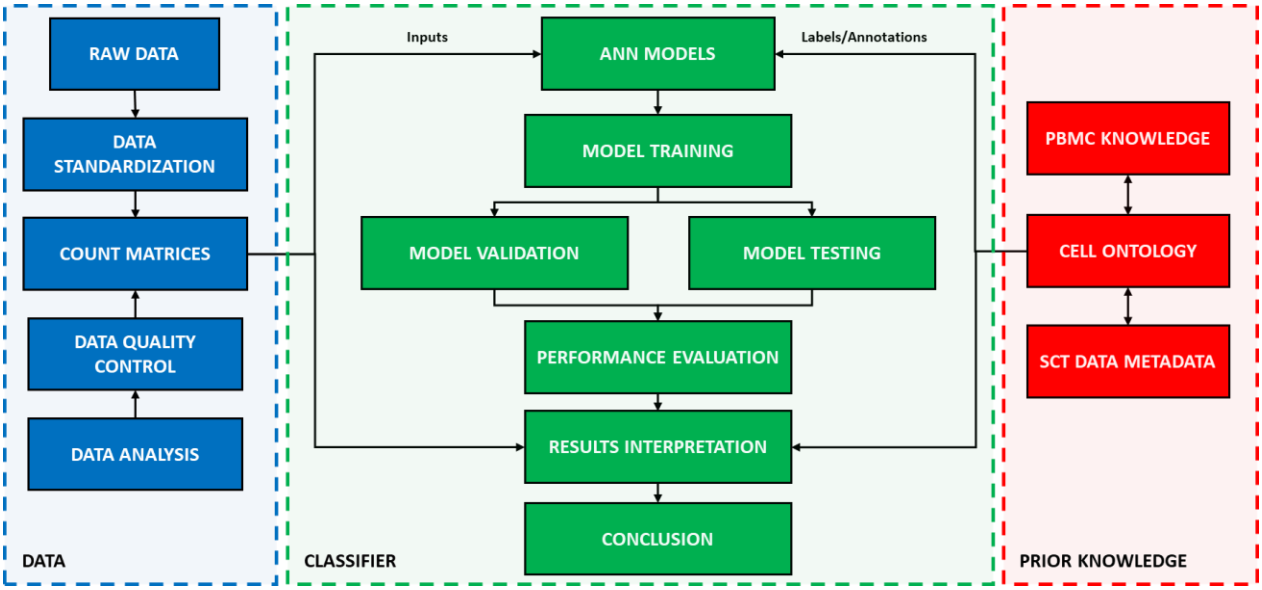


Figure 3. The technology roadmap for overall design of this project.

The datasets are collected, standardized, and stored clearly with metadata. The statistical analysis has been done with specific gene expression profiles. The data structure and distribution have been visualized to support classification procedure. Multi-dimensional PBMC-SCT cell ontology has

been developed based on PBMC prior knowledge and involved metadata. Supervised ML model ANN has been trained with quality-controlled training sets. Model validation has been done with internal cross-validation and external validation. Model testing has been done with expert-annotated, qualified testing sets. Performance assessment metrics have been used to evaluate the classification results. During incremental learning process, ANN model performance in each cyclical experiment has been recorded and assessed with certain metrics. The result of ANN classification can reflect data representativeness, data effect, PBMC-SCT ontology, and biological explanation. The system demonstrates to have good accuracy and good robustness on the generalization across multisource SCT datasets for further practical utilization.

1.5 Contribution of Thesis

CONTRIBUTION	
DATA	a) Collected and filtered independent 10x SCT data files from multiple sources. b) Made the reference gene list based on different genome versions. c) Standardized SCT data files with the reference gene list. d) Converted SCT data into different formats for various uses. e) Demonstrated a workflow of collecting, cleaning, standardizing, and converting SCT data.
METADATA	a) Made metadata for standardized SCT data files. b) The experimental information and descriptive statistical properties have been analyzed for each data set. c) Made a template for building metadata and statistical analysis.
CELL ONTOLOGY	a) Designed multi-dimensional ontology for single cell classification. b) Described PBMC-SCT cell ontology. c) Described properties of each dimension/subdimension in the ontology.

EXPERIMENT DESIGN AND MACHINE LEARNING	a) Designed training and testing experiments based on standardized SCT data. b) Proved the concept of single cell classification using SCT data and supervised ML ANN models (with overall accuracy 89.4%). c) Demonstrated internal cross-validation and external validation (with qualified testing sets). d) Performed analysis of results with determined metrics. e) Explorative experiments with datasets from different sources and different quality. f) Designed incremental learning study with ANN classification model. g) Observed the effect of data source and generating protocols to PBMC SCT classification with incremental learning (accuracy 93.0%). h) Added newly collected SCT datasets into the classification system. i) Studied 5-class classification of PBMC with 56 reference datasets and incremental learning (BC, DC, MC, NK, and TC) (94.6% of overall accuracy). j) Demonstrated external cross-validation (four-supersets-swapping training and testing, evaluating performance with datasets of different sources). k) Studied the vulnerability of ANN-SCT-PBMC classification models, using 17 non-representative datasets of five groups and 17 rounds of cyclical external cross-validation experiments.
SOFTWARE	a) Mapped data files to the genome list. Data standardization. Conversion with different formats. b) Measured statistical properties for individual dataset. c) Classifier (ANN models). d) Classifier with detailed results outputs (five scores). e) Results visualization and demonstration. Performance assessment with determined metrics.

1.6 Outline of Subsequent Chapters

The first chapter (Chapter 1) introduces the research background, motivation, hypothesis, research objectives, overall study design of this thesis, as well as the main contributions of this work. **Chapter 2** is a systematic literature review of SCT analysis for PBMC classification, the review has described the significance and challenges of supervised ML vs unsupervised ML methods in SCT single-cell classification. **In Chapter 3,** it describes the general methodology used in this project, from data & data processing (including data collection and quality control, data standardization, metadata construction, and data statistical analysis), multi-dimensional SCT cell ontology (with PBMC as an example when demonstrates the cell properties dimension), to the structure of ANN model, and the performance assessment metrics. Specific research questions, involved data sets, and study design are described separately in the chapter of each study ('Materials and Methods' of Chapters 4, 5, 6, and 7). **In Chapter 4,** single cell classification with SCT data and ANN has been demonstrated and has been proved as a concept. This is the first time demonstrating single cell classification can be done by SCT data and purely supervised ML method, the overall accuracy of PBMC classification has reached 89.4%. **In Chapter 5,** an incremental learning study design has been implemented to simulate real-life situations – the effect of data accumulation, data quality, and multiple dimensions in cell ontology, to ANN classifiers. The results have shown the generalization performance of ANN on data accumulation process by time clue, involving different data sources, sampling conditions, generation protocols, and data preprocessing methods. This chapter involves a 4-class classification of PBMC, including BC, MC, NK, and TC. **Chapter 6** is an expanded verification of SCT classification using incremental learning, newly collected datasets, and external cross-validation. The BroadS2 datasets have brought the dendritic cell class into training sets. The overall accuracy of the 5-class classification has been 94.6%. This Chapter has analyzed the effect of different SCT data protocols on model performance. **In Chapter 7,** the study on the vulnerability of ANN-SCT-PBMC classifiers has been done. It explored the model's robustness, using non-representative datasets of different properties, and cyclical external cross-validation among four data sources. The results of each study have been written and discussed within the context of each chapter **(Chapters 4, 5, 6, and 7). Chapter 8,** summarizes the entire work and looks forward to possible future work directions. **Finally,** references and appendices have been put at the end of the thesis.

CHAPTER 2 LITERATURE REVIEW - SCT Analysis for PBMC Classification

2.1 Introduction

As the key component of the immune system, peripheral blood mononuclear cells (PBMC) has been used as important research model to understand immune regulation mechanism [66-70] and as crucial clinical indicators to reflect individual's health status [35, 71-74]. With technological innovations in methodology (as shown in Figure 4), human understanding of PBMC has ranged from the cell level (with microscope), protein level (with flow cytometry), to the transcriptome level (with transcriptome technology); from the mixture of cell populations or cell groups (with bulk RNA-seq) to individual single cells (with single-cell RNA-seq). Single-cell transcriptome (SCT) sequencing technology has made it become fact to observe the instantaneous transcription profile of each individual single cell.

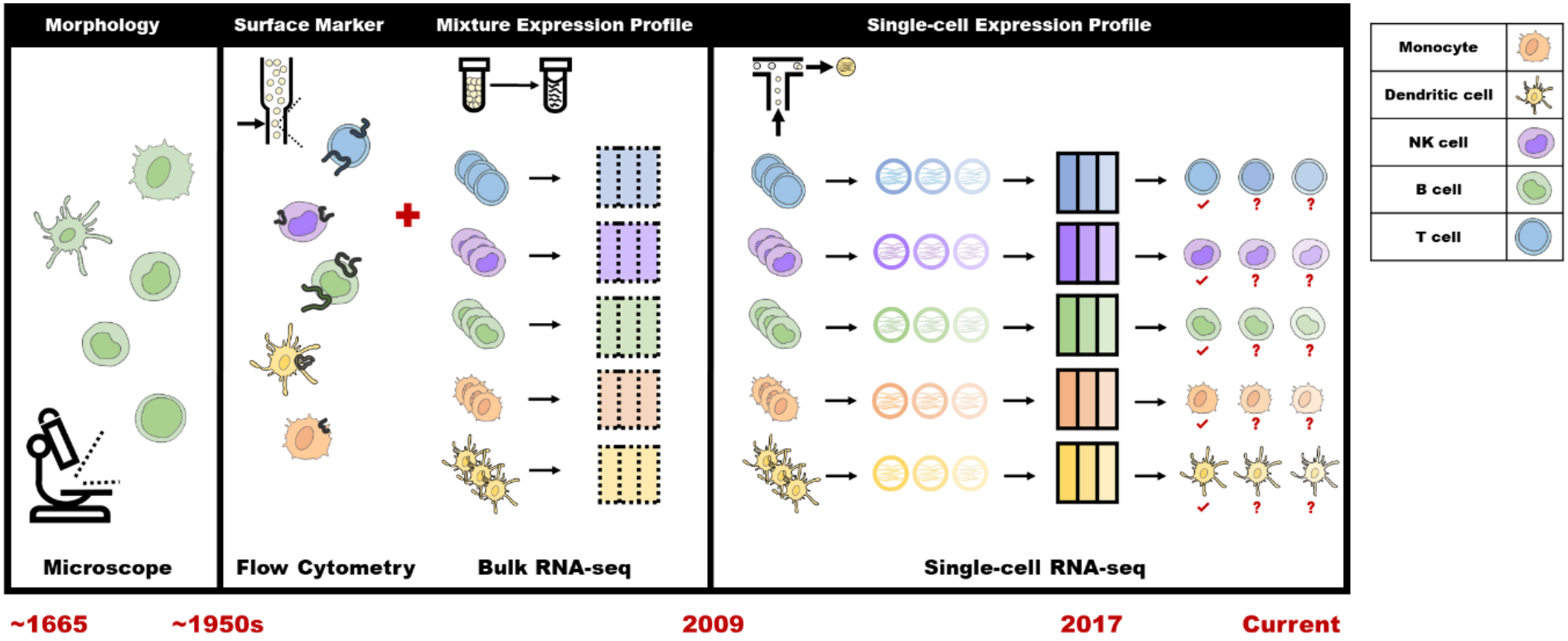


Figure 4. Illustration of technology and PBMC cell type recognition and classification strategy by time.

2.2 SCT for PBMC Study

Mainly, SCT technology in blood research has four types of applications: on medicine [75], on hematopoiesis (developmental biology), on immune cell heterogeneity (immunology), and on cell type definition (cell biology).

In medicine, SCT can help establish a transcriptome-based drug treatment monitoring for time-dependent immunotherapy (*e.g.* Ibrutinib - chronic lymphocytic leukemia (CLL)) [71, 76]; SCT can decipher human cellular immune responses (also antibody memory response) in highly detail in prophylactic vaccine development [77-79]; SCT for peripheral immune activity can help interpret the immune dynamics of severe disease processes, such as in hematological cancers [80], in infectious diseases (*e.g.* COVID-19) [81-86], and in immunodeficiency diseases (*e.g.* HIV) [87].

In hematopoiesis, SCT has challenged the classic tree model of hematopoietic lineage [88] and has provided new insights into the development model of the hematopoietic system [89, 90] and also the mechanism of blood cell differentiation in hematopoietic ageing process [91, 92].

In immune cell heterogeneity, SCT has recognized new rare cell types or intermediate cell types beyond classic well-known immune cell types. New types of dendritic cells (DC) [48, 93], monocytes (MC) [48], and CD4+ T cells (TC) [94] have been detected and profiled by SCT. In specific physiological environment or disease, the diversity of immune cell subpopulations observed by SCT can increase understanding in immune system [12, 67, 71, 75, 95].

In cell biology, the definition of "cell type" is a significant proposition [96]. After the definition by location, morphology and molecular markers [97, 98], currently SCT has redefined "cell type" on single cell transcriptome level, using SCT data – data-driven definition - SCT expression profiles [11]. With this deeper viewing angle to observe single cells' momentary states, SCT has also raised up questions on defining new PBMC cell ontology [99] and setting detailed nomenclature authentication [100] for PBMC subtypes.

2.3 Currently Used Unsupervised ML Methods and Its Limitations

The core issue for SCT in PBMC analysis is to recognize/classify/annotate PBMC cell types with SCT data. The challenges of this task stay in the natural properties of **SCT data itself** (zero-inflated, high-dimensional, large data volume, high variable sensitivity, transcriptional noise, too

informative), **the lack of generalized analysis tools, the lack of reference data set** (*i.e.*, annotated highly reproducible SCT profiles for each PBMC subset under different sampling conditions), and **the lack of uniformed experimental protocols for data integration.**

Till now, there are more than 1,000 SCT analysis tools have been developed and stored in online tools database (www.scRNA-tools.org) [101]. Many process-oriented tools and software packages have been developed, such as CellRanger [10], Seurat [102], etc. However, there is still a lack of universal tools with high repeatability in SCT analysis.

In the early stage, with the background of lacking adequate reference data sets and accurate annotations to train classifiers, **unsupervised clustering** methods and followed with empirical manual annotations have been in a dominant position in SCT data analysis. In this kind of workflow, an unsupervised algorithm is usually used to cluster a certain batch of data obtained in one study at a time, and cells with similar gene expression profiles are aggregated into discrete cell clusters. After that, algorithms (SCDE [103], DEsingle [104], SigEMD [105], SC2P [106], CRE [107], DECENT [108]) are used to recognize differentially expressed genes across cell clusters and visualization tools are deployed to check the dispersion of clusters in two-dimensional or three-dimensional data space. Significant cell identification markers are collected from literature and gene marker database to manually label cell type tags to cell clusters [48]. Automated cell label annotation tools such as, singleR [109], scmap [110], CellAssign [111], SCSA [112], scMatch [113], scCATCH [114], p-DCS [115], CellFishing.jl [116], etc. have been gradually developed to help correct the subjectivity caused by manual annotation to a certain extent.

Unsupervised clustering methods can learn single cell expression patterns and structures and classify them without annotation. In the absence of highly reproducible reference data sets and reference labels, unsupervised clustering algorithms can analyze cell heterogeneity and annotate cell types within a certain interpretable level. Also, it has made contribution to discover new heterogeneity in known cell types, to label transient cell states with featured genes, and to build hierarchical structure in single cell relationships with statistical distance.

Table 1. Unsupervised, semi-supervised, and supervised tools and packages enumerations for single cell type clustering and classification.

TYPES	METHODS	PACKAGES
UNSUPERVISED METHODS	Hierarchical clustering	ascend [117], CIDR [118], scran [119], pcaReduce [120], SCENIC [121], SINCERA [122]
	Graph-based clustering	Cell Ranger [10]
	Louvain	Seurat 1.0 [123], SCANPY [124]
	Spectral clustering	SIMLR [125]
	Density-based clustering	Monocle [126], Monocle2 [127]
	Grade of membership models	countClust [128]
	k-Medoids clustering	RaceID2 [129], RaceID3 [130]
	k-Means clustering	RaceID [131], SAIC [132], scVDMC [133]
	Consensus clustering (k-Means + Hierarchical clustering)	SC3 [56]
	Model-based clustering	TSCAN [134]
	Aggregated clustering methods	SAFE [135]
SEMI-SUPERVISED METHODS	Weighted feature genes	SCINA [136], LAmbDA [137], scANVI [138], scNym [139]
	Graph convolutional networks	scGCN [140]
SUPERVISED METHODS	Supervised hierarchical clustering	RCA [141]
	Generalized linear model classifier	Garnett [142]

	Artificial neural network (ANN)	ACTINN [143], CHETAH [144], SuperCT [145], Zhong, et al. [146]
	Support vector machine (SVM)	scPred [147], scHPL [148]
	Random Forest (RF)	SingleCellNet [149], HieRFIT [150]
	k-nearest neighbors (KNN)	SNN-Cliq [151], scClassify [152], GapClust [153]

However, unsupervised clustering methods have met its challenges and limitations in SCT analysis.

a) Lose genetic information in data preprocessing for clustering.

Clustering methods usually require proper dimension reduction methods to "project" SCT data from high dimension to lower dimension, in this process, large amount of genetic information on heterogeneity might be lost. Also, the related quality control, normalization, data correction, and feature selection methods along with this process do not benefit to preserve the integrity of genetic information. These methods have made efforts on eliminate technical variables or noises in SCT data, but they have also taken risk to remove the real biological heterogeneity information. The parameters and cutoff thresholds in these data preprocessing steps can affect the further clustering and classification performance.

b) The reusability of unsupervised clustering methods is not satisfied.

Unsupervised clustering methods for SCT analysis is one study at a time. The model developed for one data set does not generalize to other data sets. Different clustering algorithms and working flows have been applied for different independent SCT studies. The clustering results and labeling results of one same clustering tool can be various across different SCT data with diverse data sources. This is caused by the high variable sensitivity of SCT data itself and the limitations of unsupervised clustering tools. There are many reviews and testing studies for tools in clustering methods in SCT [154, 155], but so far, there is barely a unified conclusion on a generalized analysis protocol and solid widely accepted parameter settings. Most of the time, conclusions on clustering tools' accuracy, robustness, efficiency and the thresholds, parameters thereof can be made only on specific SCT data sets [4]. The lack of universality makes unsupervised clustering algorithms unable to fully integrate and utilize massive SCT data.

c) The interpretability of unsupervised clustering methods is usually not adequate.

In unsupervised learning, data instances are not labeled, and the number of classes is unknown. Unsupervised clustering methods can group single cells in visualized clusters. However, the number of clusters is artificially determined according to the degree of dispersion of cell clusters. It often happens that the number of clusters cannot be decided because the cell clusters are merged or overlapped. Clustering algorithm has challenges in interpretability and customization – clustering results might be not easy to interpret – Are cell clusters and annotations not determined arbitrarily, empirically, or in biased, in high subjectivity?

Different clustering methods and screening threshold ranges will incline to different numbers of clusters and different compositions of cell types for the same data set. At the same time, small cell clusters may not be recognized due to the limitation of the algorithm's pattern recognition resolution. Those may contain more detailed, rare, or specific cell subtypes in a deeper classification level.

Second, clustering analysis tools require that the distribution of analyzed data conform to the established statistical hypothesis. As known, SCT data is not in a typical normal distribution. After dimensionality reduction projection, it is necessary to determine whether SCT data meets the reasonableness of the hypothesis of the clustering algorithm. This helps the interpretability of unsupervised clustering analysis tools.

Third, unsupervised clustering has low sensitivity to high-dimensional SCT data. Even after dimensionality reduction and other preprocessing steps, technical errors/variables/noises caused by batch effects may affect the clustering of cell sample points more than true differences in cell transcriptome levels (i.e., cells from the same experimental source may be more likely to aggregate than cells of the same type). In addition, cell subtypes that are similar in developmental lineages cannot be accurately separated, as they have similar gene expression profiles.

These factors above can confuse the analysis and interpretation of clustering results, leading to low classification accuracy of unsupervised clustering methods.

d) The cell type marker information used in the annotation is not comprehensive.

The annotation of cell types in unsupervised clustering analysis is labor-intensive in nature and relies heavily on the analyst's knowledge and perception of cell markers, which may lead to inconsistent analysis results. At the same time, manual annotation is not suitable for large data sets. In the actual operation, the specific expression genes of the cell cluster may not match the typical marker genes of the typical cell type. At this time, the cell cluster cannot be assigned to the known

cell type. Similar cell types can share same typical markers, and some cell types may not have known typical markers.

2.4 PBMC SCT Analysis with Cell Marker

Currently, online database such as CellMarker [156], CellMatch [114], DICE [157, 158], Human Protein Atlas (http://www.proteinatlas.org) [159, 160] can support with peripheral immune cell markers in PBMC SCT analysis. Most of the cell marker information used comes from bulk-RNAseq analysis results, and many marker databases focus on the use of CD marker (cluster of differentiation marker) to type peripheral blood immune cells. It is undeniable that this type of classification criterion has formed a mature, detailed and quite accurate classification system that can be used as an authoritative reference for PBMC classification. However, it should be noted that CD marker is a cell typing standard focusing on cell surface molecules based on technologies such as flow cytometry and FACS. The transcription profile observed by SCT technology is the transient transcription level inside the cell. Deduction, identification, and determination of SCT cell types (that are based on cell transcript expression profiles) through molecules expressed on the cell surface [97], it has a certain interpretability, but there is also a huge risk of rationality.

In the current stage, at the subcellular level, endogenous cell markers (molecular markers within the cell structure, such as microRNA (miRNA) and protein) has been considered as promising SCT cell type markers. The combined use of cell surface molecular markers and endogenous markers has not been effectively deployed in the classification of SCT data.

Latest, the collection of currently known high-quality and repeatable SCT data set annotation results and the construction of a more comprehensive, unified, integrated cell annotation platform (http://celltype.info) has been carried out in multiple global single-cell research projects [161, 162].

2.5 Supervised ML in SCT Classification and Its Challenges

As a result of the constant generation of a significant number of high-quality SCT data and the rapid development of commercial single-cell sequencing platforms (e.g. 10x Genomics [10]), the number of reference data sets for single-cell classification has continued to expand. The semi-supervised and supervised learning analysis tools for SCT have been gradually developed (as shown in Table 1). The increase in publications in SCT and PBMC-SCT research fields has been demonstrated with the line chart in Figure 5.

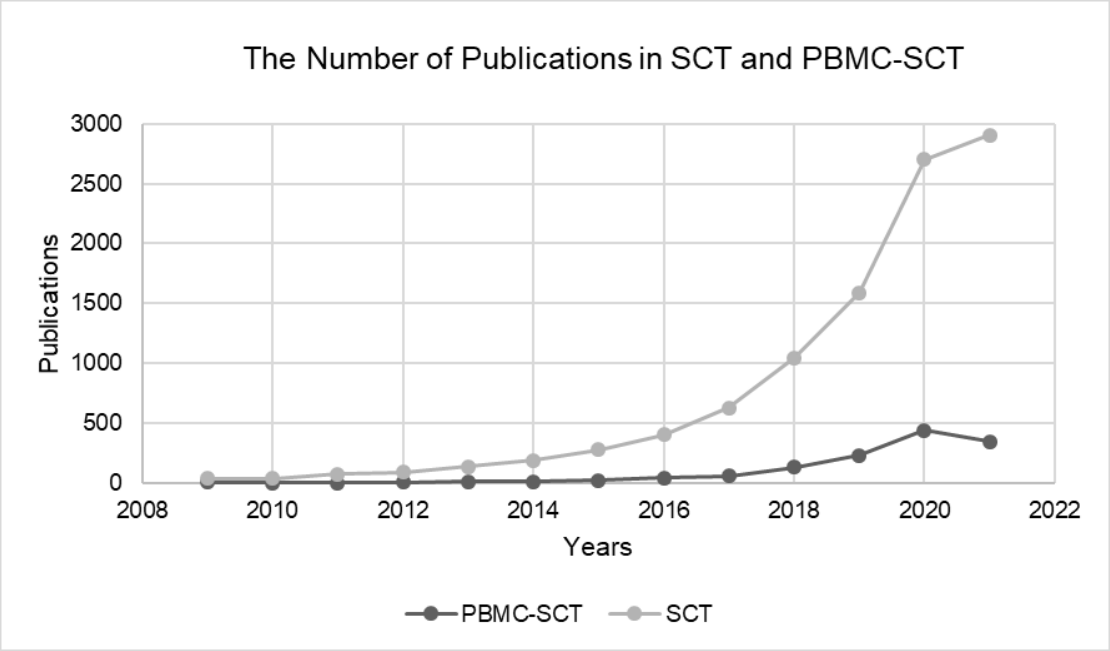


Figure 5. The increase of publications in SCT and PBMC-SCT research area by years. Data source from PubMed (pubmed.ncbi.nlm.nih.gov, NIH) with search query: for PBMC-SCT: "(single-cell transcriptomics OR single-cell RNA sequencing OR scRNA-seq) AND (peripheral blood or PBMC or circulating immune cell)"; for SCT: "(single-cell transcriptomics OR single-cell RNA sequencing OR scRNA-seq)". (The time point of data collection for this figure is 2021/09/12.)

Supervised learning classification techniques have been impressively applied to data classification, examples are network-based learning algorithms (artificial neural sanetwork (ANN), support vector machine (SVM)), and instance-based learning algorithm (*k*-nearest neighbor (*k*NN)), etc.

Supervised machine learning uses reference data sets and reference cell type labels as training data. Through learning, the supervised machine learning algorithm can accurately and effectively classify the cells of testing set, and score the confidence of the given label. Supervised machine learning is expected to effectively learn, recognize and classify SCT data expression patterns with high dimensions (~20,000 to ~30,000 feature dimensions).

a) Can handle and classify SCT data pattern.

Supervised classification methods can effectively make up for the deficiencies of unsupervised machine learning. Its advantage exists in that it can directly learn the expression pattern of the cell type from the large amount of reference data (training set) and perform reliable pattern recognition on the testing set through statistical inference algorithms. Supervised classification models such

as artificial neural networks are capable of coping with the complexity of SCT data (high-dimensional, sparse, high variable sensitivity, transcriptional noise). It can identify the unique expression patterns of specific cell types from highly variable and highly complex SCT data, and define and classify a certain cell group with the distribution of transcripts with ~20,000 to ~30,000 feature dimensions.

b) Generalization.

Supervised learning can generalize on multi-source SCT data. For SCT in PBMC classification [146], it can generalize both on sorting and non-sorting PBMC sample conditions, it can eliminate batch effect and technical variables in SCT data to a certain extent. A well-trained supervised classification model has the ability to handle with newly upcoming SCT data with various data sources.

c) Fit to large amount of SCT data.

At the same time, the huge amount of SCT data is a reasonable application of supervised learning, and the huge training set base can increase the classification accuracy of supervised learning. Supervised learning can cooperate and integrate the existing SCT data sets to maximize the utilization of SCT data resources.

However, the convinced performance of supervised classification methods has a strong dependence on the reference data set.

a) The quality of reference data.

It requires high-quality example data as training set for classification algorithm learning and building a satisfied classification model, and fitting the model to new testing set with interpretable classification results. This strictly requires a high degree of accuracy and repeatability of the training data set and its annotation labels. Low quality and contaminated training set can bring irrelevant confounding information to classification model and lead to unreliable classification results.

b) The lack of reference data on specific research samples.

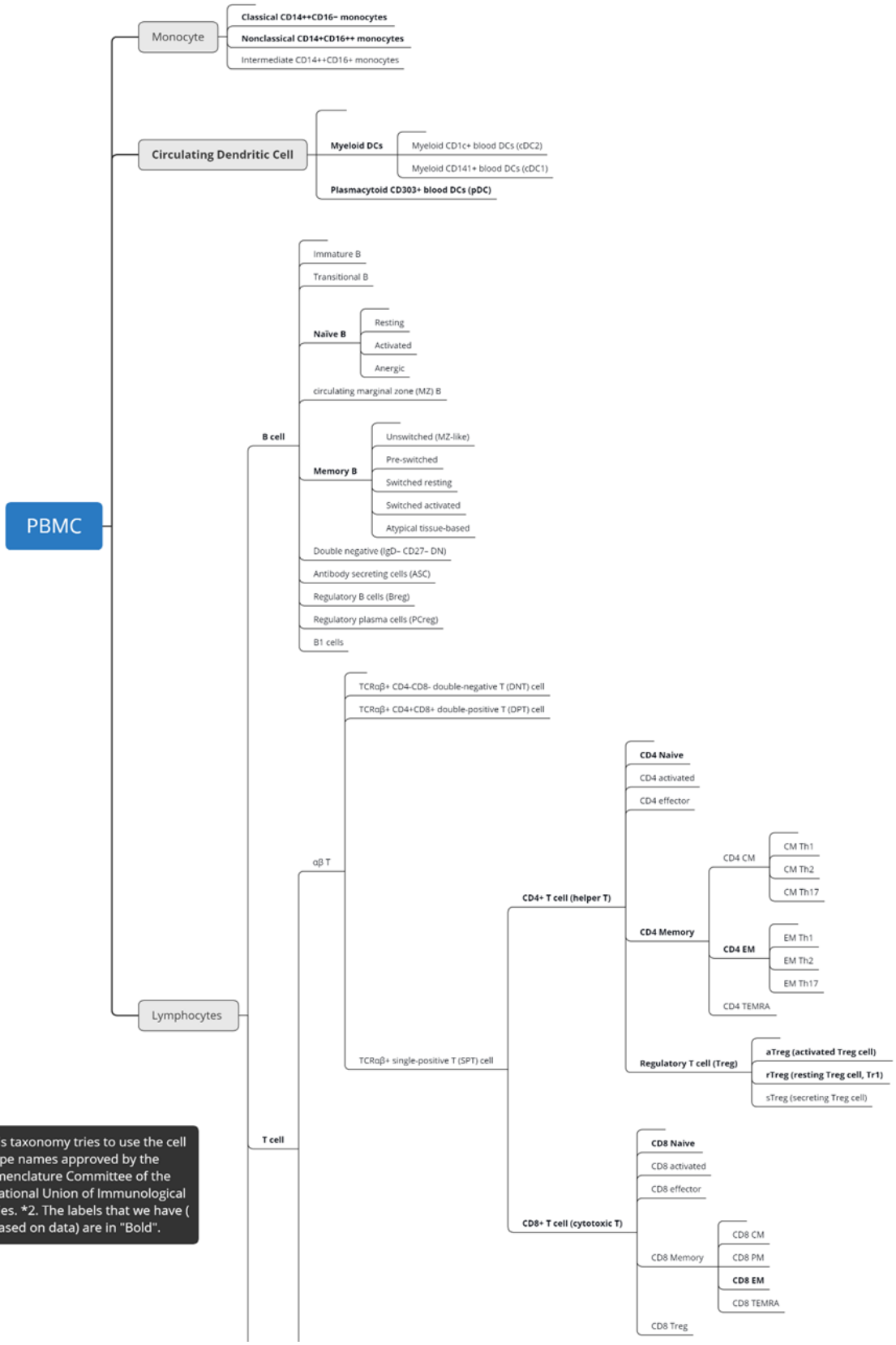
The number of SCT data sets has continued to grow exponentially, but the source of its sample tissues has become scattered for different research purposes. So far, there are SCT data sets on tissues such as liver, heart, kidney, brain, whole blood, etc., but there are few SCT reference data

sets for a specific studied cell population. That leads to a shortage of training and forming an effective classification model for a specific aim.

Moreover, due to the limitation of cell separation technology, most of the sample collection is a mixture of a certain organ and tissue, rather than a specific single cell type or cell group. This leads to a lack of sufficient training sets for a single cell type for supervised learning. For example, as far as a classification study of PBMC [146] is concerned, for healthy human peripheral blood, a total of 58 high-quality, effective and reproducible SCT data sets of PBMC subtypes has been collected from January 2017 to April 2020. In the process of collecting data sets, we have found that a large number of sample sources are whole blood or PBMC mixture, but few samples are of a single cell type with cell separation (such as pure T cells, Monocytes, or B cells samples). Among the few purified PBMC samples, most of them come from research focusing on a specific disease. Their samples are collected from patient donors with disease. There are very few data sets on PBMC of healthy human donors.

Reference data sets on certain research samples need to be generated and integrated for building satisfied SCT classification model. The following (Figure 6) is a dendrogram for PBMC ontology. It generally represents the relationships among significant PBMC cell types and subtypes. However, only those cell types highlighted in bold have accordingly SCT profiles, other cell types they are still waiting for upcoming profiles in SCT resolution. In fact, there are over hundreds of PBMC subtypes [163] have been found by previous bulk-RNAseq for a complete PBMC ontology. However, there is still no standardized SCT profiles for these subtypes. The classes and relationships among these subtypes are not clarified yet. To build a detailed SCT-PBMC ontology, the SCT profiles and hierarchical relationships for these subtypes need to be determined using SCT technology and SCT data analysis tools.

Without detailed SCT-PBMC ontology and specific SCT-PBMC subtype data, a detailed classification model with PBMC subtypes cannot be fully constructed. Currently, only five-class classification models have been constructed for PBMC SCT classification [146].


PBMC
Monocyte
Classical CD14++CD16− monocytes
Nonclassical CD14+CD16++ monocytes
Intermediate CD14++CD16+ monocytes
Circulating Dendritic Cell
Myeloid DCs
Myeloid CD1c+ blood DCs (cDC2)
Myeloid CD141+ blood DCs (cDC1)
Plasmacytoid CD303+ blood DCs (pDC)
Lymphocytes
B cell
Immature B
Transitional B
Naïve B
Resting
Activated
Anergic
circulating marginal zone (MZ) B
Memory B
Unswitched (MZ-like)
Pre-switched
Switched resting
Switched activated
Atypical tissue-based
Double negative (IgD− CD27− DN)
Antibody secreting cells (ASC)
Regulatory B cells (Breg)
Regulatory plasma cells (PCreg)
B1 cells
T cell
αβ T
TCRαβ+ CD4-CD8- double-negative T (DNT) cell
TCRαβ+ CD4+CD8+ double-positive T (DPT) cell
TCRαβ+ single-positive T (SPT) cell
CD4+ T cell (helper T)
CD4 Naive
CD4 activated
CD4 effector
CD4 Memory
CD4 CM
CM Th1
CM Th2
CM Th17
CD4 EM
EM Th1
EM Th2
EM Th17
CD4 TEMRA
Regulatory T cell (Treg)
aTreg (activated Treg cell)
rTreg (resting Treg cell, Tr1)
sTreg (secreting Treg cell)
CD8+ T cell (cytotoxic T)
CD8 Naive
CD8 activated
CD8 effector
CD8 Memory
CD8 CM
CD8 PM
CD8 EM
CD8 TEMRA
CD8 Treg
*1. This taxonomy tries to use the cell type names approved by the Nomenclature Committee of the International Union of Immunological Societies. *2. The labels that we have (based on data) are in "Bold".

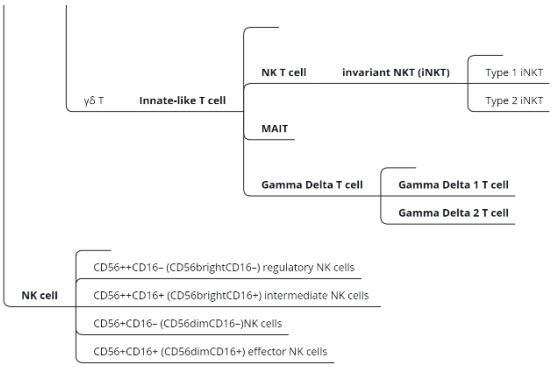


Figure 6. Organized PBMC ontology taxonomy.

c) The lack of understanding to new cell sub-class/sub-state found with SCT.

The other limitation of supervised machine learning in SCT analysis, currently, is from the incompleteness of existing cell ontology/taxonomy in multi-dimensions. Other than subtypes found in previous bulk-studies, SCT has found new intermediate/sub-subtypes along with different cell states. Supervised classification model needs more SCT data in sub-sub class (intermediates or subtypes) and in different sample conditions (as for healthy PBMC, e.g. activated, memory, or effector memory cell states) to interpret the classification results.

For example, the above PBMC ontology is mainly built based on knowledge from literature and bulk-RNAseq database. It has missed hundreds of PBMC subtype classes, those new sub-subtypes/cell states decoded by SCT technology [163].

The lack in well-defined classes for newly found sub-subtypes/cell states, that can lead to the misclassification between the two subtypes/cell states that are very close to each other on similarity (e.g. Classical CD14++CD16- Monocytes, Nonclassical CD14+CD16++ Monocytes, and Intermediate CD14++CD16+ Monocytes).

Classification model requires to learn sub-subtypes' SCT expression profiles – those are in the next/deeper classification level. These sub-subtypes have not been found in previous technologies, but they have been observed in SCT resolution [48]. The shortage in profiles and class definition (forming an entity in current PBMC ontology) for these subtypes have made 2%~3% misclassification in PBMC SCT classification [146].

Latest, the SCT project Human Cell Atlas (HCA) has been making efforts on clarifying cell types and ontologies for SCT analysis [164, 165].

At the same time, as the PBMC ontology has being amended, revised, and updated, the confirmation and clarification of the cell type nomenclature should comply with unified standards. This helps to eliminate the confusion or ambiguity of cell types, and helps to establish a more precise and rigorous classification system for cell types.

d) Supervised learning requires strict standardized operation protocols (SOPs) in SCT data generation.

Supervised learning methods can deal with the batch effect brought by different experimental protocols, different chemical agencies, and different data pre-processing protocols to a certain degree, but it still has around a 1%~2% misclassification rate [146] coming from lack of unified,

strict SOPs. It has been found that with the increase in the number of highly reproducible training sets, the classification accuracy of the supervised learning model can come over the batch effect and converge to a certain level.

SCT data with strict SOPs is helpful in performance of supervised machine learning in detailed SCT classification. Unified SOPs for SCT data generation is expected to promote real SCT application in clinical precision medicine in the future.

2.6 Combination of Supervised and Unsupervised ML in SCT

The latest SCT data classification should consider the combination of unsupervised clustering and supervised classification methods - that can better improve the accuracy of cell classification and recognition. The analysis results of the two types of methods can be referred to each other.

Supervised classification can verify the results of unsupervised analysis of cell clusters. Supervised classification uses high-quality reference data sets and high-accuracy reference labels to ensure the classification results more reliable and interpretable. This can make up for the subjectivity in unsupervised clustering analysis.

While at the same time, for new, unknown intermediate cell types or rare cell types found in supervised classification (those have not been successfully classified), unsupervised analysis can be used to help annotate new cell subtypes and identify their specific differential expression genes. This helps to update and refine the existing cell ontology and enrich the classification layers of supervised classification.

Supervised and unsupervised learning can help each other, promote each other, and help enrich and deepen the understanding of existing cell types.

2.7 Current Challenges in SCT Classification Analysis

So far, the enormous efforts have been made both in supervised and unsupervised learning tools for SCT analysis. Currently, there are some challenges that still hinder the large-scale integrated application analysis of SCT data.

a) Technique deficiency.

The first essential challenge comes from the technique deficiency hiding in SCT technology itself. As known, SCT technology can capture the most ~70% transcriptome information in a single cell, still ~30% genetic information can be missed in SCT profiles.

This leads to the confusion understanding of “zero” value in SCT profiles. There are two possible reasons for the inference of zero value, one is the real zero expression of the transcript (i.e., the transcript does not exist), and the other is that the transcript is not captured (i.e. dropout event) due to the shallow sequencing depth. About 90% of the values in the SCT expression profile matrices are zero values. Too many zero values cause raw SCT data to present an irregular zero-inflated negative binomial distribution instead of a normal distribution in statistics. The reasonable judgment and understanding of the zero value have always been one of the main challenges of SCT data quality control and classification analysis.

Another example of the noise caused by technical factors is doublets and triplets. In the single cell capture process, two or three cells and one gel bead are wrapped together by one oil droplet, that will cause "the cell" (a collection of two or three cells) to show an exceptionally specific high-level RNA expression. The understanding and processing of doublets and triplets also brings challenges to SCT analysis.

Next generation technology is expected to solve these technical confounding factors and decode SCT profiles of single cells in more comprehensive and more accurate level.

b) Challenges from the understanding of single cell biology.

Due to technological advancement, SCT has given humans an unprecedented opportunity to observe the transcriptional profile of a transient snapshot of a single cell. However, even if the interference of all technical factors is hypothetically ignored (assuming that there is no dropout, no batch effect, SCT data sets are all high-quality, reproducible, generated with a strictly unified protocol), the super microscopic level of SCT observation also makes humans lack sufficient existing knowledge to explain the captured biological phenomenon of single cells.

The transient expression state of single PBMCs is coordinated by a variety of factors, including cell differentiation state (from naïve, immature, to mature), cell proliferation state (different cell cycle stages - G1, S, G2, M - circulating immune cells keep the ability of mitogenesis and proliferation [166, 167]), cell activation state (antibody activated or cytokine activated, memory

or effector memory state), and cell transcriptional bursting state [168] (The transcription activity in the cell is not continuous, but pulsed. At a certain stage or moment, the high-intensity expression will be ushered in. SCT will capture a snapshot of transient expression, that may be at the peak period or the trough period of expression.). At these moments of different states, the same "type" of cells will have a great difference in expression, and this will bring great influence and confusion to distinguish different cell types with SCT data.

This has also triggered a redefinition of cell types in the single-cell era: Should cells be classified according to all the observable transient states of cells? Or just focus on the stable cell state over a period of time? How should we clarify and define "a type" of cells [11, 96]? If consider all the SCT transient states of cells, the PBMC ontology can add hundreds of new subtypes. How should the single-cell PBMC ontology be reconstructed using multiple dimensions?

c) Establishing global unified, standardized, strict SOPs and systematic workflows for SCT.

SCT profiles to a same cell type can be influenced by the protocols both in experimental sequencing, data preprocessing, and data computational analysis.

For example, in PBMC single cell sequencing process, with the difference in cell separation methods, sampling conditions (fresh PBMC or frozen-thawed PBMC), sampling temperatures, storage time, sequencing protocols (10x, smart-seq2, smart-seq3), chemical reagents (chemical v2, v3 for 10x); the PBMC frequency, cell viability, cell transcription level can be affected, and digital SCT profiles can show different results.

The similar in SCT data preprocessing and computational analysis processes, different parameter and thresholds selection will make differences in final SCT profiles.

Formulating and establishing global unified, standardized SOPs (from SCT sequencing to raw data standardization, data analysis) for SCT benefits to global SCT data concordance, integration, and comprehensive utilization. Many experts have put forward opinions and suggestions [169-171] on the formation of a standardized and unified strict SOPs for PBMC SCT sampling, storage, sequencing, and data analysis workflow.

Globalized SCT projects such as HCA [42] and other single-cell genomics consortiums have raised strict standardization requirements [172] and systematic workflow models [13] for SCT data generation.

Large-scale global integrated generation and analysis of SCT data is the only way to go, that not

only meets its requirements as biomedical Big Data, but also meets the needs of supervised machine learning. A large number of highly standardized reference data sets help to achieve the repeatability and comparability of SCT data. That can maximize the elimination of the influence of technical factors, help set the quality control threshold used to limit technical noise, help determine individual differences, and help determine the possibility of a certain disease risk.

d) Lack of reference data and reference annotation for detailed cell subtypes.

As has been discussed above, the lack of reference data sets and detailed labels for specific aims has largely limited the current PBMC SCT analysis. There is currently a huge demand gap for high-quality annotation and high reproducibility SCT data of PBMC subtypes under different sampling conditions.

It should be noted that it may not be possible to generate authentic and reliable labels for all cell subtypes [97]. Due to the inherent defects of SCT technology, the comprehensive multi-dimensional information of several dynamic cell subtypes may not be captured. The lack of true labels and reference data sets for all cell subtypes [173] is an essential obstacle for machine learning in SCT analysis.

e) Lack of generalized analysis tools.

There is still a need for generalized analysis tools with high robustness, accuracy, and scalability, to respond to the massive exponential growth of single-cell data.

At the same time, there is a need for uniformity in the programming language and input data format of the analysis software. The current analysis software is mainly written in R language and Python, and the input formats are various across different software. Achieving flexible conversion between different analysis software and input objects is the key to user-friendliness.

f) Establishing unified SCT data storage and transfer platform.

The big data [19] nature of single-cell data requires it to form a global data storage and coordination platform. High-quality, repeatable and standardized SCT cell profiles should be stored in integrated data platform.

Bulk-RNAseq has made examples in blood /immune cell reference databases, such as NovershternHematopoieticData [174], DatabaseImmuneCellExpressionData (DICE) [157], and MonacoImmuneData [175].

In SCT, the HCA project has formed a data coordination model (data.humancellatlas.org) [162, 165] for reference data sets. The Atlas of Blood Cells (ABC) project has made reference data sets for 7551 human blood cells of 21 healthy donors with SCT [176]. A global systematic data platform is required to be designed for these treasurable PBMC-SCT data sets.

2.8 Future Prospects for PBMC-SCT Classification

Despite the enormous challenges of biological cognition and computational analysis, we can see the broad prospects of PBMC-SCT data for clinical precision medicine.

With the exponential growth of PBMC-SCT data, and the continuous expansion and combined use of unsupervised clustering and supervised machine learning in the SCT field, accurate and robust recognition of the expression patterns of PBMC-SCT profiles will become a reality.

The complexity and diversity of the massive PBMC-SCT profiles implies the judgment of individual health, disease, age, or clinical drug treatment effects. A sufficient number of standardized PBMC-SCT data sets with accurate class labels, can be used as the basis for predicting genetic phenotypes and decision making of clinical diseases.

Large-scale integrated PBMC-SCT data analysis is expected to become an essential category in electronic health record (EHR) [173] system, and hopes to become an information-based disease prevention and monitoring method, for blood diseases, cancer [177], immune diseases, and infectious diseases in the future.

CHAPTER 3 GENERAL METHODOLOGY

3.1 Data

3.1.1 Data collection & data processing

The 10x SCT data sets collected for this project study mainly have four sources, these are 10x Genomics Demonstration Data, GEO database, BroadS1 study and BroadS2 study.

The 10x Genomics Demonstration Data is the database supported and maintained by 10x Genomics company, that represents the high-quality PBMC data sets generated with standardized 10x experimental protocol. BroadS1 and BroadS2 studies are accomplished by Broad Institute with specific and clear cell type annotation for PBMC sample data sets. They are considered as precisely high-quality data sets and can be used as training and testing data sets for the supervised machine learning PBMC classification system.

For GEO database, the 10x SCT sequencing data of relevant articles published by 13th July 2019 were searched using keywords - “single cell” AND “10x” in GEO (Gene Expression Omnibus) Database of NCBI (National Center of Biotechnology Information, https://www.ncbi.nlm.nih.gov/). In total, 595 10x SCT data sets of *Homo Sapiens* in GEO database have been collected. Among these collected 595 GEO 10x SCT data sets, specific data sets using PBMC as experimental samples have been selected, stored, and annotated one by one.

Raw data (matrix.mtx, barcodes.tsv, genes.tsv) of Study BroadS1, Study BroadS2, GEO data sets, and 10x Genomics Demonstration data have been downloaded, collected, filtered, and stored. Data sets with specific annotation of one cell type of PBMC and generated by PBMC sampling from healthy donors have been selected as the training sets with specific classes for building the classification system initially. The data sets annotated with PBMC mixture sample have been stored and prepared to use for the following experiment purpose to test the robustness of the classification model system.

Collected data sets involves different publication date, different sample source, and different experimental condition in collected data sets. Raw data usually contains one gene list file, one barcode sequence file and one gene expression matrix file for each study. Data files corresponding to their data source and study source have been organized and stored in local data repository and the backup files have been made in different local storage terminals. Data backup is also uploaded

to the cloud data storage server.

3.1.2 General metadata construction

Metadata contains useful traceability information of involved data sets, that consists of two main parts, one is experimental metadata, one is statistical metadata. The experimental metadata includes the study description, study number, sample name, experimental condition, cell type, technology platform of each data set collected, that gives background experimental information of each study. The statistical metadata includes data distribution and basic statistical properties of each data set, that helps to understand the difference and data structure in each count matrix.

INDEX	RIES	ID: ACCESSION	No S	GENOME	SAMPLE	ORGANISM	No S	PLATFORM	DATA FOR	TYPE	STUD	AUTH	DESCRIPTION	C	Date	Short description	Tissue	Cell typ	Receptor	
212	35	200109037 GSE109037	11	GRCh38	GSM2928382	Homo sapiens	1	GPL21290	MTX, TSV	Expre	10x G	Subr	AdultHuman_17-5: Adult testis, Homo sapie	11/6/2018	OK	Steady-state Spermatogenic cells				
213	35	200109037 GSE109037	11	GRCh38	GSM2928383	Homo sapiens	1	GPL21290	MTX, TSV	Expre	10x G	Subr	AdultHuman-Spermatocytes_17-6: Adult te	11/6/2018	OK	Sta-Put enriched Spermatocytes				
214	35	200109037 GSE109037	11	GRCh38	GSM2928384	Homo sapiens	1	GPL21290	MTX, TSV	Expre	10x G	Subr	AdultHuman-Spermatids_17-6: Adult testis	11/6/2018	OK	Sta-Put enriched Spermatocytes				
216	35	200109037 GSE109037	11	GRCh38		Homo sapiens	1	GPL21290	MTX, TSV	Expre	10x G	Subr	AdultHuman-Spermatocytes_17-11: Adult t	11/6/2018	OK	Sta-Put enriched Spermatocytes				
217	35	200109037 GSE109037	11	GRCh38		Homo sapiens	1	GPL21290	MTX, TSV	Expre	10x G	Subr	AdultHuman-Spermatids_17-11: Adult testis	11/6/2018	OK	Sta-Put enriched Spermatocytes				
276	42	200110686 GSE110686	2	GRCh38	GSM3011853	Homo sapiens	1	GPL16791	MTX, TSV	Expre	Single	Subr	Primary TNBC infiltrating T cells case 1; Prima	6/24/2018	OK	T cells - tumor infiltrating T cells TNBC	breast cancer	T cells	CD3+	
277	42	200110686 GSE110686	2	GRCh38	GSM3011854	Homo sapiens	1	GPL16791	MTX, TSV	Expre	Single	Subr	Primary TNBC infiltrating T cells case 2; Prima	6/24/2018	OK	T cells - tumor infiltrating T cells TNBC	breast cancer	T cells	CD3+	
286	46	200111014 GSE111014	12	GRCh38	GSM3020393	Homo sapiens	1	GPL20301	MTX, TSV	Expre	Chror	Subr	PBMCs_CLL patient after 0 days of ibrutinib t	2/22/2019	OK	PBMC - CLL patient 1 no treatment	PBMCs			
287	46	200111014 GSE111014	12	GRCh38	GSM3020394	Homo sapiens	1	GPL20301	MTX, TSV	Expre	Chror	Subr	PBMCs_CLL patient after 120 days of ibrutinil	2/22/2019	OK	PBMC - CLL patient 1 after 120d treatment	PBMCs			
288	46	200111014 GSE111014	12	GRCh38	GSM3020395	Homo sapiens	1	GPL20301	MTX, TSV	Expre	Chror	Subr	PBMCs_CLL patient after 0 days of ibrutinib t	2/22/2019	OK	PBMC - CLL patient 5 no treatment	PBMCs			
289	46	200111014 GSE111014	12	GRCh38	GSM3020396	Homo sapiens	1	GPL20301	MTX, TSV	Expre	Chror	Subr	PBMCs_CLL patient after 150 days of ibrutinil	2/22/2019	OK	PBMC - CLL patient 5 after 150d treatment	PBMCs			
290	46	200111014 GSE111014	12	GRCh38	GSM3020397	Homo sapiens	1	GPL20301	MTX, TSV	Expre	Chror	Subr	PBMCs_CLL patient after 30 days of ibrutinib	2/22/2019	OK	PBMC - CLL patient 5 after 30d treatment	PBMCs			
291	46	200111014 GSE111014	12	GRCh38	GSM3020398	Homo sapiens	1	GPL20301	MTX, TSV	Expre	Chror	Subr	PBMCs_CLL patient after 0 days of ibrutinib t	2/22/2019	OK	PBMC - CLL patient 6 no treatment	PBMCs			
292	46	200111014 GSE111014	12	GRCh38	GSM3020399	Homo sapiens	1	GPL20301	MTX, TSV	Expre	Chror	Subr	PBMCs_CLL patient after 120 days of ibrutinil	2/22/2019	OK	PBMC - CLL patient 6 after 120d treatment	PBMCs			
293	46	200111014 GSE111014	12	GRCh38	GSM3020400	Homo sapiens	1	GPL20301	MTX, TSV	Expre	Chror	Subr	PBMCs_CLL patient after 280 days of ibrutinil	2/22/2019	OK	PBMC - CLL patient 6 after 280d treatment	PBMCs			
294	46	200111014 GSE111014	12	GRCh38	GSM3020401	Homo sapiens	1	GPL20301	MTX, TSV	Expre	Chror	Subr	PBMCs_CLL patient after 30 days of ibrutinib	2/22/2019	OK	PBMC - CLL patient 6 after 30d treatment	PBMCs			
295	46	200111014 GSE111014	12	GRCh38	GSM3020402	Homo sapiens	1	GPL20301	MTX, TSV	Expre	Chror	Subr	PBMCs_CLL patient after 0 days of ibrutinib t	2/22/2019	OK	PBMC - CLL patient 8 no treatment	PBMCs			
296	46	200111014 GSE111014	12	GRCh38	GSM3020403	Homo sapiens	1	GPL20301	MTX, TSV	Expre	Chror	Subr	PBMCs_CLL patient after 120 days of ibrutinil	2/22/2019	OK	PBMC - CLL patient 8 after 120d treatment	PBMCs			
297	46	200111014 GSE111014	12	GRCh38	GSM3020404	Homo sapiens	1	GPL20301	MTX, TSV	Expre	Chror	Subr	PBMCs_CLL patient after 30 days of ibrutinib	2/22/2019	OK	PBMC - CLL patient 8 after 30d treatment	PBMCs			
300	49	200111108 GSE111108	1	GRCh38	GSM3022245	Homo sapiens	1	GPL18573	CSV	Expre	Single	Subr	Mixture of H2228, H1975 and HCC827 huma	2/28/2018	OK					
316	55	200111912 GSE111912	2	GRCh38	GSM3044891	Homo sapiens	1	GPL20301	TXT	Expre	The c	Subr	scRNA-Seq by 10X, replicate 1: lymphoblast	6/14/2018	OK					
317	55	200111912 GSE111912	2	GRCh38	GSM3044892	Homo sapiens	1	GPL20301	TXT	Expre	The c	Subr	scRNA-Seq by 10X, replicate 2: lymphoblast	6/14/2018	OK					
326	59	200112845 GSE112845	9	GRCh38	GSM3087619	Homo sapiens	1	GPL18573	MTX, TSV	Expre	PBMC	Subr	DTM-X_PBMC_live; whole blood; Homo sapi	7/25/2018	OK	PBMC - DTM-X_PBMC_live	healthy whole blc	PBMCs		
327	59	200112845 GSE112845	9	GRCh38	GSM3087622	Homo sapiens	1	GPL18573	MTX, TSV	Expre	PBMC	Subr	DTM-X_PBMC_methanol-3H+SSC; whole blc	7/25/2018	OK	PBMC - DTM-X_PBMC_methanol-3H+SSC, su	healthy whole blc	PBMCs		
328	59	200112845 GSE112845	9	GRCh38	GSM3087624	Homo sapiens	1	GPL18573	MTX, TSV	Expre	PBMC	Subr	DTM-X_PBMC_methanol-3W+SSC; whole bl	7/25/2018	OK	PBMC - DTM-X_PBMC_methanol-3W+SSC, su	healthy whole blc	PBMCs		
329	59	200112845 GSE112845	9	GRCh38	GSM3087626	Homo sapiens	1	GPL18573	MTX, TSV	Expre	PBMC	Subr	DTM-Y_PBMC_methanol-3W+SSC; whole bl	7/25/2018	OK	PBMC - DTM-Y_PBMC_methanol-3W+SSC, su	healthy whole blc	PBMCs		
330	59	200112845 GSE112845	9	GRCh38	GSM3087628	Homo sapiens	1	GPL18573	MTX, TSV	Expre	PBMC	Subr	CD8+_live; whole blood; Homo sapiens; isola	7/25/2018	OK	T cells - live, whole blood, CD8+	whole blood	T cells	CD8+	ir
331	59	200112845 GSE112845	9	GRCh38	GSM3087629	Homo sapiens	1	GPL18573	MTX, TSV	Expre	PBMC	Subr	CD8+_methanol_SSC; whole blood; Homo sa	7/25/2018	OK	T cells - whole blood, suspension D, CD8+	whole blood	T cells	CD8+	ir
332	59	200112845 GSE112845	9	GRCh38	GSM3087631	Homo sapiens	1	GPL18573	MTX, TSV	Expre	PBMC	Subr	KLM1_live; Pancreatic cancer cell line; Homo	7/25/2018	OK					
333	59	200112845 GSE112845	9	GRCh38	GSM3087633	Homo sapiens	1	GPL18573	MTX, TSV	Expre	PBMC	Subr	KLM1_methanol_PBS; Pancreatic cancer cell	7/25/2018	OK					
334	59	200112845 GSE112845	9	GRCh38	GSM3087634	Homo sapiens	1	GPL18573	MTX, TSV	Expre	PBMC	Subr	KLM1_methanol_SSC; Pancreatic cancer cell	7/25/2018	OK					
339	62	200113036 GSE113036	1	GRCh38	GSM3136660	Homo sapiens	1	GPL16791	CSV, MTX	Expre	Perio	Subr	10X_ASCL1_SOX2_DORSO_d14: 10X_ASCL1_	6/5/2018	OK					
346	64	200113196 GSE113196	4	GRCh38	GSM3099646	Homo sapiens	1	GPL20301	TXT	Expre	Single	Subr	Individual 4: Mammoplasty Reduction; Hom	4/17/2018	OK					
347	64	200113196 GSE113196	4	GRCh38	GSM3099647	Homo sapiens	1	GPL20301	TXT	Expre	Single	Subr	Individual 5: Mammoplasty Reduction; Hom	4/17/2018	OK					
348	64	200113196 GSE113196	4	GRCh38	GSM3099848	Homo sapiens	1	GPL20301	TXT	Expre	Single	Subr	Individual 6: Mammoplasty Reduction; Hom	4/17/2018	OK					
349	64	200113196 GSE113196	4	GRCh38	GSM3099849	Homo sapiens	1	GPL20301	TXT	Expre	Single	Subr	Individual 7: Mammoplasty Reduction; Hom	4/17/2018	OK					
352	66	200113660 GSE113660	1	GRCh38	GSM3110765	Homo sapiens	1	GPL20301	MTX, TSV	Expre	UMI-	Subr	Rh41 scRNA-seq; Rh41 single cell; Homo sa	4/30/2018	OK					

Figure 7. The components of metadata involving over 600 10x SCT files.

The above figure is the structure of the metadata of this project involving over 600 data sets, that shows the component and modality of the designed metadata chart form. The aggregated data annotation of the 10x SCT studies has been arrayed into the metadata chart form, that is designed with “INDEX”, “SERIES”, “ACCESSION”, “GENOME”, “ORGANISM”, “DESCRIPTION”, “SAMPLE TYPE” etc. as the captions of each column in metadata. The metadata is sorted by “ACCESSION”, that is the number name of series (e.g. GSE119561). ACCESSION is arranged in order from small to large, from top to bottom. This is very important to the follow-up work, because it has been found that many related data sets have very similar series numbers.

Only 10x SCT technology relevant research is included in metadata, other research with other single cell transcriptomics technologies (e.g. Drop-seq, SMART-seq, inDrop, etc.) of the same super series is not involved in. Sample number (e.g. GSM3377671) is unique for each 10x study

in GEO database. Data sets collected from other sources, such as 10x Genomics Demonstration, BroadS1 study, BroadS2 study, they have their own unique sample indexes. In this study, data sets from different sources have been renamed and reorganized based on the research purpose.

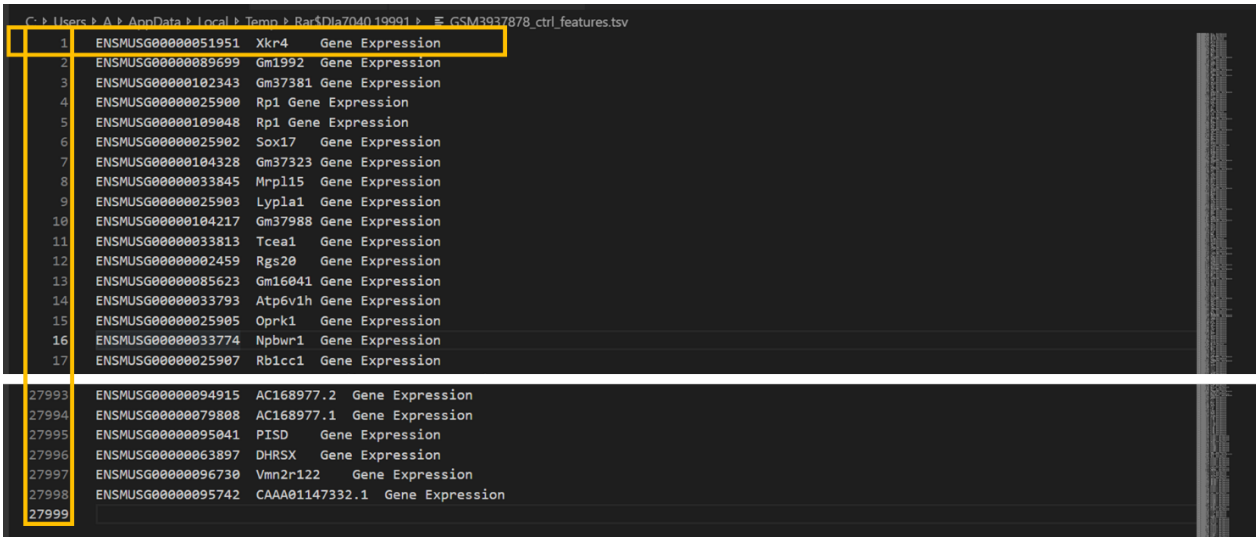
The comprehensive metadata has over 600 data sets mapped with their own studies, the description of each study is involved in the metadata and some of them has specific additional comments. The metadata has detailed annotation for each specific data set. It has described the sample cell type, health status of the donors, experimental conditions, experimental protocols, data upstream analysis protocols, and other important information of each 10x SCT data set for the further experimental design of the supervised machine learning PBMC classification system construction.

3.1.3 Data selection and study quality control

During 10x SCT data collection process, the good quality of collected data sets has been checked and ensured for the further following pre-processing steps and classification steps. For example, in GEO SCT data sets collecting process, the studies which are not related but filtered out by GEO database browser with the key words are excluded (e.g. 10X Hank's salt solution). Another example is that series with inconsistent study description are excluded out as well.

3.1.4 Common genome assembly built

Genome assembly is the gene name database comprises the names and IDs of all known genes so far, it is used as available annotation tracks. Different genome version is used in different studies. The alteration of genomic versions and the lack of uniform naming standards have led to complex confusion. One gene name can have several different probes name, it is not comparable between two different genomes of one same organism. Quality control has been done to exclude studies only supply gene name list without probes or only have probes list without gene name list. One probe can correspond to different gene names (synonym or alias). NCBI, ENSEMBL and UCSC are genome databases and genome browsers retrieving genomic information. The number of probes in genome assembly are regularly updated. Genome assembly has Ensembl Gene ID (e.g. ENSG00000210049) and Gene Name (e.g. MT-TF).



```
C: ▸ Users ▸ A ▸ AppData ▸ Local ▸ Temp ▸ Rar$Dla7040.19991 ▸ GSM3937878_ctrl_features.tsv
1     ENSMUSG00000051951  Xkr4     Gene Expression
2     ENSMUSG00000089699  Gm1992   Gene Expression
3     ENSMUSG00000102343  Gm37381  Gene Expression
4     ENSMUSG00000025900  Rp1 Gene Expression
5     ENSMUSG00000109048  Rp1 Gene Expression
6     ENSMUSG00000025902  Sox17    Gene Expression
7     ENSMUSG00000104328  Gm37323  Gene Expression
8     ENSMUSG00000033845  Mrpl15   Gene Expression
9     ENSMUSG00000025903  Lypla1   Gene Expression
10    ENSMUSG00000104217  Gm37988  Gene Expression
11    ENSMUSG00000033813  Tcea1    Gene Expression
12    ENSMUSG00000002459  Rgs20    Gene Expression
13    ENSMUSG00000085623  Gm16041  Gene Expression
14    ENSMUSG00000033793  Atp6v1h  Gene Expression
15    ENSMUSG00000025905  Oprk1    Gene Expression
16    ENSMUSG00000033774  Npbwr1   Gene Expression
17    ENSMUSG00000025907  Rb1cc1   Gene Expression

27993 ENSMUSG00000094915  AC168977.2   Gene Expression
27994 ENSMUSG00000079808  AC168977.1   Gene Expression
27995 ENSMUSG00000095041  PISD     Gene Expression
27996 ENSMUSG00000063897  DHRSX    Gene Expression
27997 ENSMUSG00000096730  Vmn2r122     Gene Expression
27998 ENSMUSG00000095742  CAAA01147332.1   Gene Expression
27999
```

Figure 8. An example of genome assembly (GSM3937878).

We used the current version (.tsv) in ENSEMBL genome browser as reference. In our study, genome builds have been selected of different samples in different series from collected data. They have been compared and merged to a dictionary of reference genome assembly, it is named as "common list", with probes mapping to genes.

1	ALL PROBES HUMAN						
2							
3	PROBES	hg19	GRCh37	GRCh38	Ensembl_GRCh38.p12_rel94	GSM3717979	
4	ENSG00000117533	hg19_VAMP4	grch37_VAMP4	grch38_VAMP4	#VAMP4	#VAMP4	in all
5	ENSG00000228915				#OR7E128P		Ensembl_GRCh38.p12_rel94
6	ENSG00000248222	hg19_CTB-174D11.1	grch37_CTB-174D11.1	grch38_CTB-174D11.1	#AC011389.1	#AC011389.1	in all
7	ENSG00000236230	hg19_RP11-400N13.1	grch37_RP11-400N13.1	grch38_RP11-400N13	#AL356108.1	#AL356108.1	in all
8	ENSG00000236596				#AC092568.1		Ensembl_GRCh38.p12_rel94
9	ENSG00000233029	hg19_RP11-439A17.9	grch37_RP11-439A17.9	grch38_RP11-439A17	#AC244453.2	#AC244453.2	in all
10	ENSG00000162636	hg19_FAM102B	grch37_FAM102B	grch38_FAM102B	#FAM102B	#FAM102B	in all
11	ENSG00000261714				#AC105137.1		Ensembl_GRCh38.p12_rel94
60566	ENSG00000101871	hg19_MID1	grch37_MID1	grch38_MID1	#MID1	#MID1	in all
60567	ENSG00000196517	hg19_SLC6A9	grch37_SLC6A9	grch38_SLC6A9	#SLC6A9	#SLC6A9	in all
60568	ENSG00000092439	hg19_TRPM7	grch37_TRPM7	grch38_TRPM7	#TRPM7	#TRPM7	in all
60569	ENSG00000221840	hg19_OR4A5	grch37_OR4A5	grch38_OR4A5	#OR4A5	#OR4A5	in all
60570	ENSG00000284387				#MIR24-2		Ensembl_GRCh38.p12_rel94
60571	ENSG00000085733	hg19_CTTN	grch37_CTTN	grch38_CTTN	#CTTN	#CTTN	in all
60572	ENSG00000168140	hg19_VASN	grch37_VASN	grch38_VASN	#VASN	#VASN	in all
60573	ENSG00000258631	hg19_RP11-739G5.1	grch37_RP11-739G5.1	grch38_RP11-739G5.1	#AC110023.1	#AC110023.1	in all
60574							

Figure 9. Comparison across different genome version.

Correction has been made when the genomes adopted in several studies show the wrong data format, the decimal point in probe, the space keys, confused/mixed genome version and the incorrect naming. Corrected and cleaned genome file is saved with format ".txt" or ".tsv" instead of ".csv", in case of Excel date format confusion. Genome files supplied in ".H5" file format are converted to ".csv" format. The cleaned and merged version of genome assembly is used as reference for follow-up machine learning section.

Table 2. Components and the number of gene probes in common list and full list of *Homo Sapiens*.

Organism	Component	Common list probes number	Full list probes number	Note
Homo Sapiens	"grch37_1_GSM3073089_GSE112570.tsv" "grch38_1_GSE117403.tsv" "grch38_2_GSM3375767_GSE119506.tsv" "grch38_3_GSM3478791_GSE122703.tsv" "grch38_4_GSM3543618_GSE124703.tsv" "grch38_5_GSM3813936_GSE131685.tsv" "hg19_2_GSM3430548_GSE121267.tsv" "hg19_3_GSM3635372_GSE127471.tsv" "hg19_4_GSM2897333_GSE108394.tsv" ... ... (special genome)	30698	60570	"hg19_1_GSM2867931_GSE106245.tsv" "hg19_5_GSM3143601_GSE114530.tsv" Deleted. (Decimal point, date format error, version error.)

In this study, a common gene list across collected SCT data of *Homo Sapiens* has been prepared as a mapping library for count matrix standardization. In human common gene list, there are totally 30698 gene names with corresponding gene probes, they are features in PBMC SCT classification.

3.1.5 Data filtering, conversion, and standardization

Raw data sets have been filtered, decompressed, converted into standardized file formats. Data files of super series studies have been split up.

Figure 10. Data files collected and cleaned.

Data cleaning and filtering has been done by the exclusion of null data, that can be caused by the count of empty droplets during SCT experiment procedure – the cell capture rate is zero at this situation. For example, in the raw data of the study sample GSM3258348, the cell barcode is ~700,000, but the actual gene expression is only ~26,000, that means it calculates lots of empty

gene expression, so we filtered output the actual meaningful data by removing the null data in each matrix of raw data file.

```
%%MatrixMarket matrix coordinate integer general
%
30698 1760 2466878
2 1 1
5 1 2
20 1 1
53 1 5
96 1 1
115 1 1
116 1 1
158 1 1
208 1 2
244 1 2
250 1 20
268 1 1
277 1 1
305 1 1
310 1 1
319 1 1
```

Figure 11. MTX file needs to be converted to CSV file for visualization.

Raw data of different formats (e.g. .h5, .csv, .tsv, .txt, .mtx) with different genome versions have been converted into CSV file (.csv), with cell barcodes/cell numbers as the horizontal heading, the standard 30,698 gene features as the vertical heading, and gene expression values as digital matrix. The produced CSV file was converted into four standard file formats - .h5, .csv, .npz, .mtx (tsv), those used as common, unified and standardized output format for various purpose of use, such as file transfer, visualization, and statistical calculation.

	A	B	C	D	E	F	LHL	LHM	LHN	LHO	LHP	LHQ	LHR	LHS	LHT	LHU	LHV	LHW	LHX	LHY	LHZ
1		1	2	3	4		8331	8332	8333	8334	8335	8336	8337	8338	8339	8340	8341	8342	8343	8344	
2	hg38_ZFP41	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	hg38_ASF1A	0	0	0	2		0	0	0	0	0	2	0	0	0	0	0	0	3	0	
4	hg38_TMEM39A	0	0	0	0		0	0	0	0	0	2	0	0	0	0	0	3	0	0	
5	hg38_OR2M7	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	hg38_CCT7	0	0	0	0		0	3	0	3	0	3	4	0	0	0	0	0	3	0	
7	hg38_DDX60	0	0	0	0		0	0	0	0	3	0	0	0	0	0	0	3	0	0	
8	hg38_EPOR	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9	hg38_NECAP2	0	0	0	0		0	0	0	3	0	0	0	0	0	3	0	0	0	0	
10	hg38_RP11-63N8.3	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	hg38_EDNRA	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
12	hg38_CTD-2034I21.1	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	hg38_RP4-735C1.4	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	hg38_SLC43A1	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30681	hg38_CDC20	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30682	hg38_DNAJB9	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30683	hg38_OSTM1	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30684	hg38_SPATA8-AS1	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30685	hg38_ARMC2	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30686	hg38_AADAC	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30687	hg38_KRTAP6-1	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30688	hg38_LCNL1	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30689	hg38_LINC00535	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30690	hg38_GS1-24F4.2	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30691	hg38_TRIM31-AS1	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30692	hg38_PHB	4	0	0	0		0	3	0	0	0	0	4	0	0	0	0	0	0	3	
30693	hg38_DNHD1	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30694	hg38_CTB-129P6.11	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30695	hg38_TPH2	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30696	hg38_LINC00626	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30697	hg38_LRRC29	0	0	0	2		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30698	hg38_AC073333.8	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30699	hg38_ZFP36	0	0	0	0		0	3	0	3	0	0	3	0	0	0	0	0	0	0	
30700																					
30701																					

BS1_T_cells_std_83

Figure 12. An example of a standardized count matrix (30,698 features).

Data standardization has mapped the original digital matrix in raw data set to reference common

gene list (30,698 features). Those gene probes in common list that don't have expression in cells have been filled up with zeros. Original gene probes in raw data that are not involved in reference list have been filtered out.

3.1.6 PBMC data selection and properties analysis

Among the collected SCT data sets, PBMC data sets with 'blood' as sample sources have been sorted out for following studies. There are 9 data sets of 10x Genomics Demonstration, 28 data sets of GEO database, 5 data sets of BroadS1 study and 31 data sets of BroadS2 study.

3.1.6.1 PBMC data metadata

The experimental information (experiment platform, experimental conditions, sample sources, etc.) and statistical information (cell number, etc.) of PBMC data sets have been described in metadata.

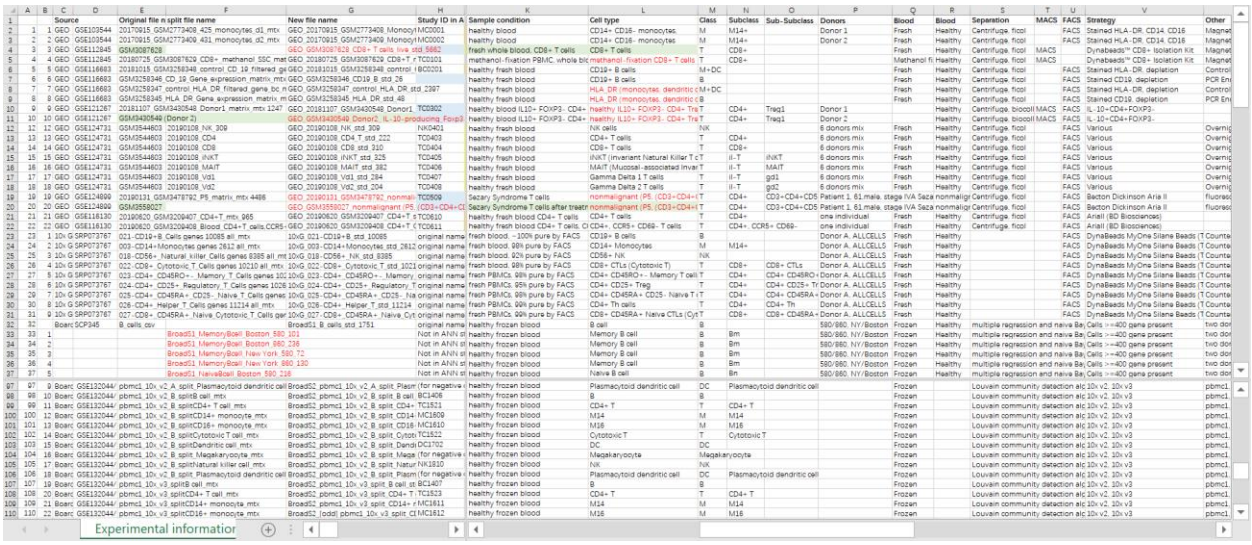


Figure 13. The experimental metadata and statistical metadata for involved PBMC data sets.

In PBMC metadata, original file names have been renamed with the index number of the study. PBMC data sets have been arrayed according to index, data source, original file name, new file name, publication date, study ID/series number/accession number, data format, experimental platform and protocol, genome, study description, sample source, cell type, receptors, special conditions, cell ranger version (the chemical), cell sorting method, etc.

Cyclical PBMC classification experimental design can be done based on the selection of these

prepared and standardized data sets. The experimental metadata can help to explain and interpret ANN classifier behavior when it comes to multisource data sets.

3.1.6.2 Basic statistical analysis

The statistical properties of each data sets have been calculated, analyzed, stored in the statistical metadata. The statistical metadata contains information such as cell number, min value, max value, medium value, average value, sum profile, positive profile (gene expressed profile), normalized sum profile, percentile of sum and positive values, etc. for each data file matrix.

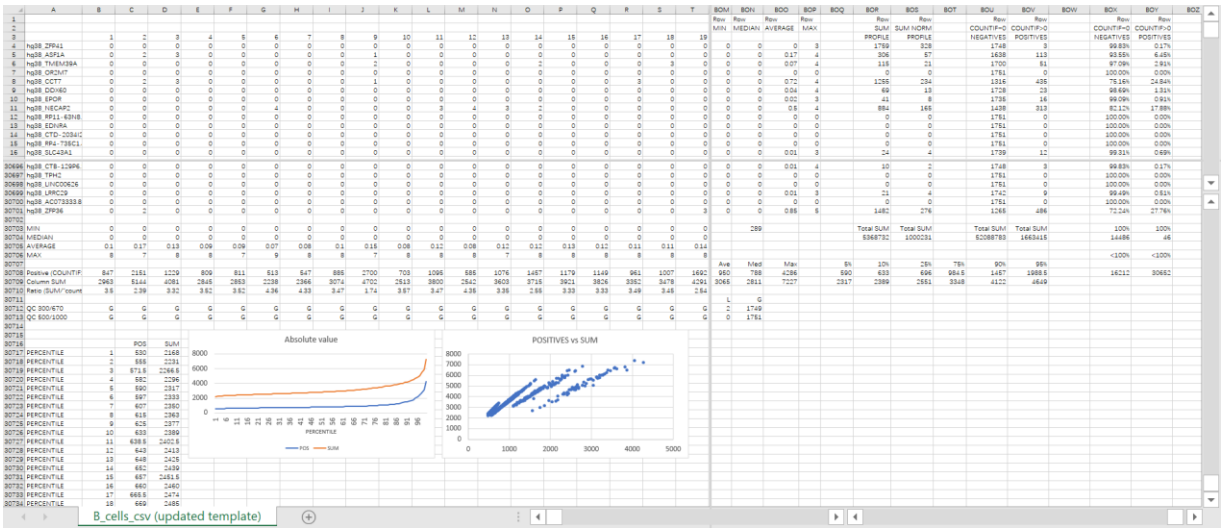
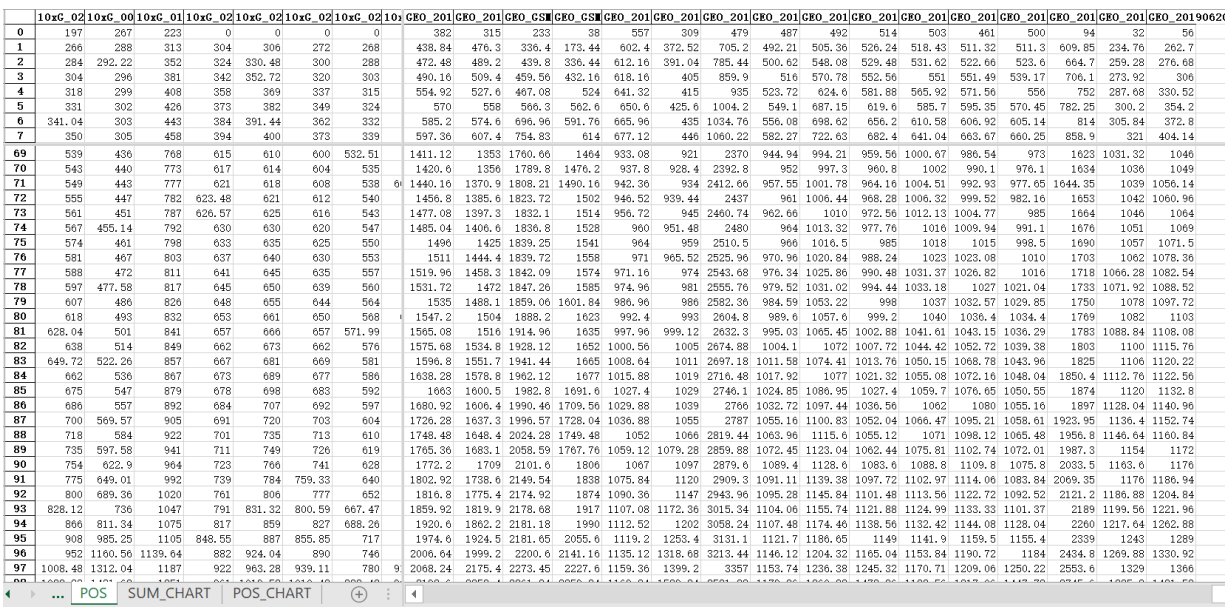


Figure 14. An example to show the statistical properties calculating procedure for one individual data set.



	10xG_02	10xG_00	10xG_01	10xG_02	10xG_02	10xG_02	10xG_02	10x	GEO_201	GEO_201	GEO_CSM	GEO_GSM	GEO_201	GEO_201	GEO_201	GEO_201	GEO_201	GEO_201	GEO_201	GEO_201	GEO_201	GEO_201	GEO_201	GEO_201	GEO_2019062
0	197	267	223	0	0	0	0		382	315	233	38	557	309	479	487	492	514	503	461	500	94	32		56
1	266	288	313	304	306	272	268		438.84	476.3	336.4	173.44	602.4	372.52	705.2	492.21	505.36	526.24	518.43	511.32	511.3	609.85	234.76		262.7
2	284	292.22	352	324	330.48	300	288		472.48	489.2	439.8	336.44	612.16	391.04	785.44	500.62	548.08	529.48	531.62	522.66	523.6	664.7	259.28		276.68
3	304	296	381	342	352.72	320	303		490.16	509.4	459.56	432.16	618.16	405	859.9	516	570.78	552.56	551	551.49	539.17	706.1	273.92		306
4	318	299	408	358	369	337	315		554.92	527.6	467.08	524	641.32	415	935	523.72	624.6	581.88	565.92	571.56	556	752	287.68		330.52
5	331	302	426	373	382	349	324		570	558	566.3	562.6	650.6	425.6	1004.2	549.1	687.15	619.6	585.7	595.35	570.45	782.25	300.2		354.2
6	341.04	303	443	384	391.44	362	332		585.2	574.6	696.96	591.76	665.96	435	1034.76	556.08	698.62	656.2	610.58	606.92	605.14	814	305.84		372.8
7	350	305	458	394	400	373	339		597.36	607.4	754.83	614	677.12	446	1060.22	582.27	722.63	682.4	641.04	663.67	660.25	858.9	321		404.14
69	539	436	768	615	610	600	532.51		1411.12	1353	1760.66	1464	933.08	921	2370	944.94	994.21	959.56	1000.67	986.54	973	1623	1031.32		1046
70	543	440	773	617	614	604	535		1420.6	1356	1789.8	1476.2	937.8	928.4	2392.8	952	997.3	960.8	1002	990.1	976.1	1634	1036		1049
71	549	443	777	621	618	608	538	6	1440.16	1370.9	1808.21	1490.16	942.36	934	2412.66	957.55	1001.78	964.16	1004.51	992.93	977.65	1644.35	1039		1056.14
72	555	447	782	623.48	621	612	540		1456.8	1385.6	1823.72	1502	946.52	939.44	2437	961	1006.44	968.28	1006.32	999.52	982.16	1653	1042		1060.96
73	561	451	787	626.57	625	616	543		1477.08	1397.3	1832.1	1514	956.72	945	2460.74	962.66	1010	972.56	1012.13	1004.77	985	1664	1046		1064
74	567	455.14	792	630	630	620	547		1485.04	1406.6	1836.8	1528	960	951.48	2480	964	1013.32	977.76	1016	1009.94	991.1	1676	1051		1069
75	574	461	798	633	635	625	550		1496	1425	1839.25	1541	964	959	2510.5	966	1016.5	985	1018	1015	998.5	1690	1057		1071.5
76	581	467	803	637	640	630	553		1511	1444.4	1839.72	1558	971	965.52	2525.96	970.96	1020.84	988.24	1023	1023.08	1010	1703	1062		1078.36
77	588	472	811	641	645	635	557		1519.96	1458.3	1842.09	1574	971.16	974	2543.68	976.34	1025.86	990.48	1031.37	1026.82	1016	1718	1066.28		1082.54
78	597	477.58	817	645	650	639	560		1531.72	1472	1847.26	1585	974.96	981	2555.76	979.52	1031.02	994.44	1033.18	1027	1021.04	1733	1071.92		1088.52
79	607	486	826	648	655	644	564		1535	1488.1	1859.06	1601.84	986.96	986	2582.36	984.59	1053.22	998	1037	1032.57	1029.85	1750	1078		1097.72
80	618	493	832	653	661	650	568		1547.2	1504	1888.2	1623	992.4	993	2604.8	989.6	1057.6	999.2	1040	1036.4	1034.4	1769	1082		1103
81	628.04	501	841	657	666	657	571.99		1565.08	1516	1914.96	1635	997.96	999.12	2632.3	995.03	1065.45	1002.88	1041.61	1043.15	1036.29	1783	1088.84		1108.08
82	638	514	849	662	673	662	576		1575.68	1534.8	1928.12	1652	1000.56	1005	2674.88	1004.1	1072	1007.72	1044.42	1052.72	1039.38	1803	1100		1115.76
83	649.72	522.26	857	667	681	669	581		1596.8	1551.7	1941.44	1665	1008.64	1011	2697.18	1011.58	1074.41	1013.76	1050.15	1068.78	1043.96	1825	1106		1120.22
84	662	536	867	673	689	677	586		1638.28	1578.8	1962.12	1677	1015.88	1019	2716.48	1017.92	1077	1021.32	1055.08	1072.16	1048.04	1850.4	1112.76		1122.56
85	675	547	879	678	698	683	592		1663	1600.5	1982.8	1691.6	1027.4	1029	2746.1	1024.85	1086.95	1027.4	1059.7	1076.65	1050.55	1874	1120		1132.8
86	686	557	892	684	707	692	597		1680.92	1606.4	1990.46	1709.56	1029.88	1039	2766	1032.72	1097.44	1036.56	1062	1080	1055.16	1897	1128.04		1140.96
87	700	569.57	905	691	720	703	604		1726.28	1637.3	1996.57	1728.04	1036.88	1055	2787	1055.16	1100.83	1052.04	1066.47	1095.21	1058.61	1923.95	1136.4		1152.74
88	718	584	922	701	735	713	610		1748.48	1648.4	2024.28	1749.48	1052	1066	2819.44	1063.96	1115.6	1055.12	1071	1098.12	1065.48	1956.8	1146.64		1160.84
89	735	597.58	941	711	749	726	619		1765.36	1683.1	2058.59	1767.76	1059.12	1079.28	2859.88	1072.45	1123.04	1062.44	1075.81	1102.74	1072.01	1987.3	1154		1172
90	754	622.9	964	723	766	741	628		1772.2	1709	2101.6	1806	1067	1097	2879.6	1089.4	1128.6	1083.6	1088.8	1109.8	1075.8	2033.5	1163.6		1176
91	775	649.01	992	739	784	759.33	640		1802.92	1738.6	2149.54	1838	1075.84	1120	2909.3	1091.11	1139.38	1097.72	1102.97	1114.06	1083.84	2069.35	1176		1186.94
92	800	689.36	1020	761	806	777	652		1816.8	1775.4	2174.92	1874	1090.36	1147	2943.96	1095.28	1145.84	1101.48	1113.56	1122.72	1092.52	2121.2	1186.88		1204.84
93	828.12	736	1047	791	831.32	800.59	667.47		1859.92	1819.9	2178.68	1917	1107.08	1172.36	3015.34	1104.06	1155.74	1121.88	1124.99	1133.33	1101.37	2189	1199.56		1221.96
94	866	811.34	1075	817	859	827	688.26		1920.6	1862.2	2181.18	1990	1112.52	1202	3058.24	1107.48	1174.46	1138.56	1132.42	1144.08	1128.04	2260	1217.64		1262.88
95	908	985.25	1105	848.55	887	855.85	717		1974.6	1924.5	2181.65	2055.6	1119.2	1253.4	3131.1	1121.7	1186.65	1149	1141.9	1159.5	1155.4	2339	1243		1289
96	952	1160.56	1139.64	882	924.04	890	746		2006.64	1999.2	2200.6	2141.16	1135.12	1318.68	3213.44	1146.12	1204.32	1165.04	1153.84	1190.72	1184	2434.8	1269.88		1330.92
97	1008.48	1312.04	1187	922	963.28	939.11	780	9	2068.24	2175.4	2273.45	2227.6	1159.36	1399.2	3357	1153.74	1236.38	1245.32	1170.71	1209.06	1250.22	2553.6	1329		1366

Figure 15. The 0-100 percentiles of positive profiles of 10x and GEO data sets as an example.

The statistical properties of each data set have been plotted into graphs for visual comparative data analysis, to figure out and contrast the difference in data structure and density distribution. The data structure and distribution represent specific gene expression profile pattern, that are crucial to ANN model performance on learning and predicting PBMC SCT data.

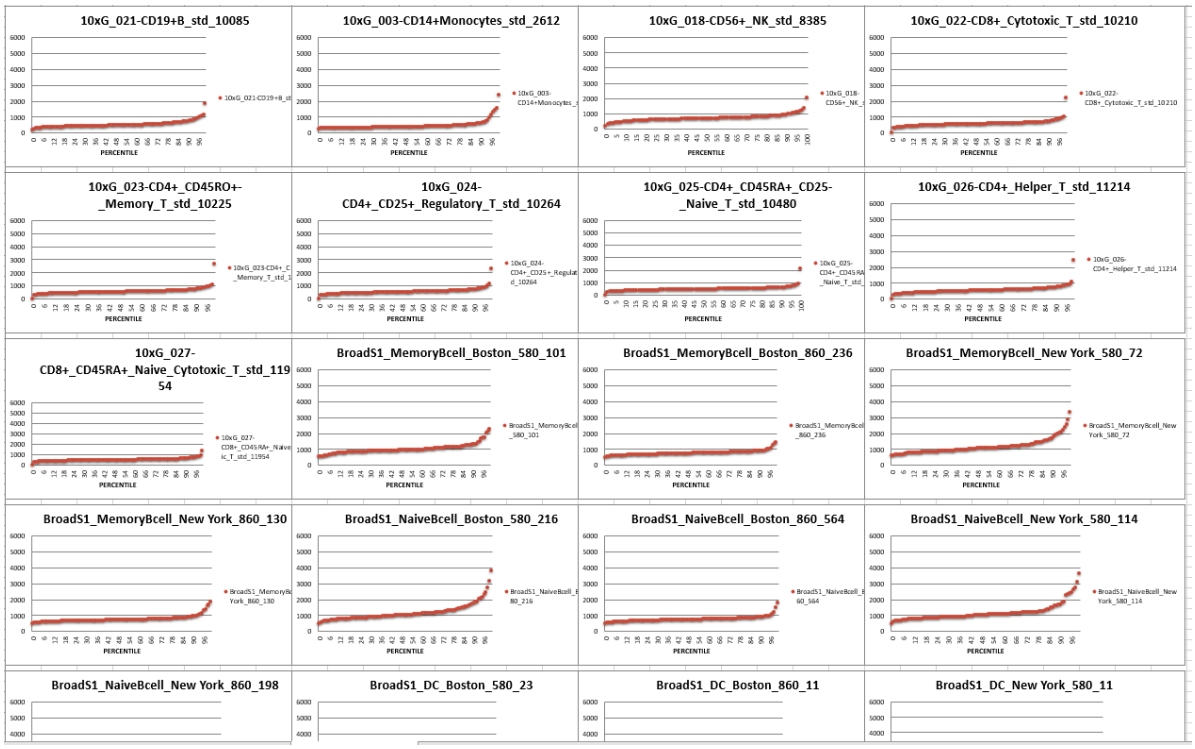


Figure 16. The scatter plots for percentiles of column positive value of each data set.

The scatter plots represent an example of statistical metadata for a data property - the percentiles of column (cell number) positive value of each data file. Through visualization using scatter plots, data distribution of each data set can be explored and analyzed on a further level.

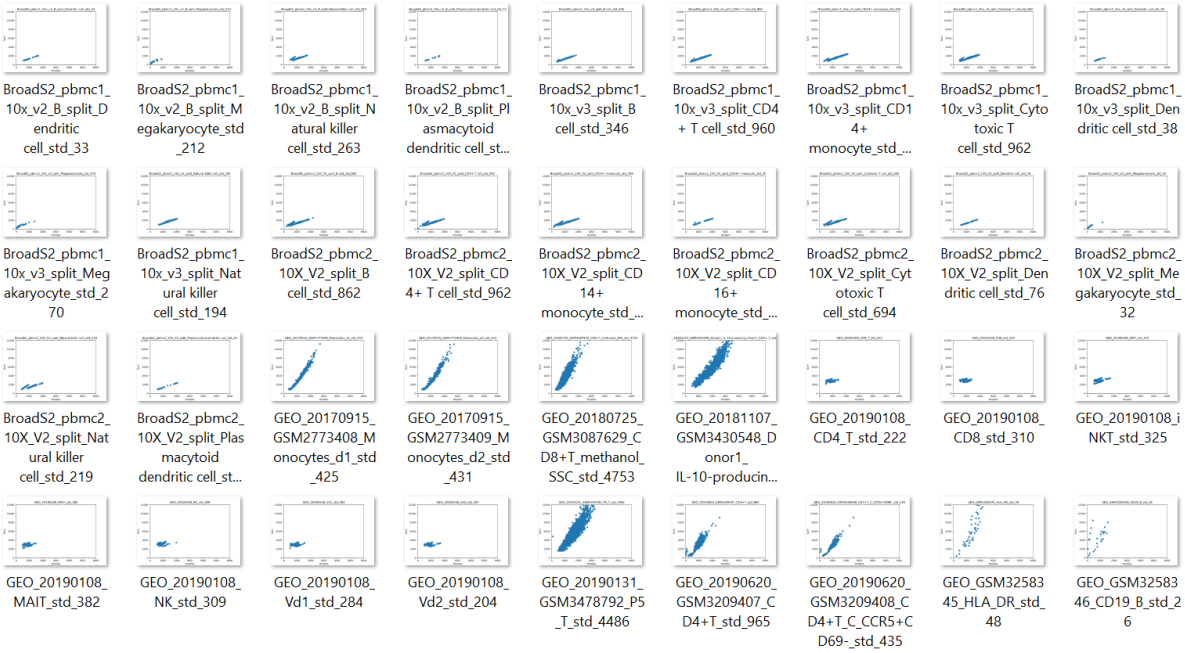


Figure 17. The scatter plots of positive values and sum values in each data set matrix.

With scatter plots of positive values and sum values, the data density and structure can be easily visualized. Based on difference in gene count thresholds, data quality control has been considered to conduct during data processing. The high expression cells can be doublets or triplets of single cells generated during 10x sequencing procedure. The low expression cells have possibility to be low-quality cell or the fragmented transcripts of single cells that should be eliminated from following supervised classification process. The differential expression analysis to SCT data sets is significant for interpreting the learning process of ANN models.

3.1.6.3 PBMC ontology metadata

A PBMC ontology has been organized based on selected PBMC SCT data, as shown in Figure 6. The ontology metadata has been organized as shown in Figure 18.

									Ad	Frequency in P	Phenotype Marker (Incomplete)	Core markers	Additional markers	Function/Pro	Ne	The labels that	Yes	GE(	10X	Bo	Ba	Background information for 5 main	Referenc
PBMC	Circulating Dendritic Cell									1~2% [1], 0.5~1%	[41]	CD3−CD19−CD20−CD14−CD56−HLA−DR+				DC	1		√	√		DCs in blood are less mature and have no	[1][2][42][4
		conventional DC (cDC)/ myeloid DC										CD3−CD19−CD20−CD14−CD56−HLA−DR+	CD1c+CD303−										[1][27]
			Myeloid CD1c+ blo							0.6% [6]	CD1c, CD11c, CD123, CD13, CD33, CD32, CD64, FcεRI, CD2, CD45RO, CD141. A small proportion also express CD14			Produce IL-12 u	mo								[1][2][3][6][
			Myeloid CD141+ bl							0.05% [6]	Clec9a, XCR1, CD11c, CD13, CD33			IL-12 and IFN-λ	mu								[1][2][3][6][
		Plasmacytoid C	D303+ blood DCs								CD303+ (BDCA-2), CD304 (BDCA-4/Neuropilin-	CD3−CD19−CD20−CD14−CD56−HLA−DR+	CD1c−	Upon pathogen		plasmocytoid d	1			√			[1][2][3][6][
	Monocyte (5~10% [1], 10-									5~10% [1], 10~30%	[39], 4~13% [41], 20.6 ± 8.91% [89]	CD3−CD19−CD20−CD14+											[27]
		Classical CD14	++CD16− monocytes							Mean: 17.7% Min:	8.34% Max: 40.1% [90]	CD3−CD19−CD20−CD14+CD16−				CD14+Monocyt	1	GSI	√	√	√		[46][27][90
		Nonclassical CD								Mean: 1.55% Min:	0.38% Max: 4.41% [90]	CD3−CD19−CD20−CD14+CD16+			pr	CD16+ Monocy	1			√	√	M16 - which comprise up to 10% of the mo	[46][27][90
		Intermediate CD																					[46]
	Lymphocytes (70 - 90% [									70 - 90% [39]													[39]
		B cell (5~10% of								5~10% of PBMC [1	CD3−CD19+					CD19+ B cell, E	1	GSI	√		√	Total B cell, Anchor marker CD45, CD sub	[11][12][13][
			Immature B																				[41]
			Transitional B								CD3−/CD19+/CD20+/CD24++/CD38++	IgD+CD27−CD38+CD24+										pre-naive	[11][38][41]
				T1B (T1/T2?)								IgD+ CD27− CD38++ CD24++	CD10++/+ IgM++ MTG	Developmental precursor									[41][47]
				T2 B (T2-MZP (MZ								IgD+ CD27− CD38++ CD24++ CD21++	CD10+ IgM++	MZ precursor									[41][47]
				T3 B								IgD+ CD27−CD38+ CD24+ CD21+	CD10+/− IgM+ MTG+	Developmental precursor or activated naïve									[41][47]
			Naïve (mature) B							Mean 5.2% Min 1.5	CD3−/CD19+/CD20+/CD27−/IgD+	CD19+, IgD+, CD27−, CD38+/−, CD24+/−				Naive B cell	1		√			express the BCR as IgM and IgD molecule	[9][10][11][4
				Resting naïve B								IgD+ CD27−CD38+ CD24+ CD21+	IgM+ MTG−	Antigen inexperienced mature cells									[47]
				Activated naive B								IgD+ CD27− CD38− CD24− CD21−	IgM+ MTG+ CD95+ CD	Precursor of short-lived PB and GC reactions									[9][10][41][4
				Anergic naive B						<2% of blood B cells, 30% of all Naïve B cells [47]		IgD+ CD27− CD38+/lo CD24+ CD21−	IgMlow/−	Hypo-responsive. Maintenance of tolerance									[47]
			Circulating marginal zone (MZ) B																				[47]
			Memory B								CD3−/CD19+/CD27+/IgD+/−	CD19+, CD27+, CD38+/−, CD24+/−				Memory B cell	1		√				[9][10][11][4
				Unswitched (MZ-like)								IgDlo CD27+ CD38+/lo CD24+ CD21+	CD1c+ IgM++	Natural memory. MZ equivalent									[41][47]
					IgM-only and IgM+IgD+ memory B cell																IgM-only B cells (IgM+IgDlow/−) and IgM+I	[52][47]
				Pre-switched							CD3−CD19+IgD+CD27+	IgD− CD27+ CD38+/lo CD24+ CD21+	IgM+	Pre-switch memory, early IgM memory. IgG memory precursor									[47][27]
				Switched							CD3−CD19+IgD−CD27+											Switched memory B cells (SM B cell)	[27][41][47
					IgG+ memory B cell																	Approximately 15~20% of peripheral blood	[52][47]
					IgA+ memory B cell																	IgA+ memory B cells account for approxim	[52][47]
				TCRαβ+ single-pc																			[41]
					CD4+ T cell (helper T					25~60% [1][39], 2	CD3+CD4+CD8−					CD4+ T cell, CD4+		GSI	√		√	CD4 T cell can be further classified into va	[11][39][69]
						Naïve CD4+ T cell	(naive differentiated helper T cel	CD4		Mean 25.9% in PB	CD3+/CD8−/CD4+/CCR7+/CD45RA+	CD45RA+ CCR7+	CD197+			T4naive, CD4+	1		√	√			[11][20][62]
						Activated CD4+ T cell					CD3+/CD8−/CD4+/CD38+/HLADR+												[66]
						Effector CD4+ T cell					CD3+/CD8−/CD4+/CCR7−/CD45RA+												[66][85]
						Memory CD4+ T cell				Mean 54.6% in PB	[62]	CD4+ CD45RO+				CD4+ CD45RO	1		√				[62][84][88
							Central memory help			Mean 40.2% in PB	CD3+/CD8−/CD4+/CCR7+/CD45RA−	CD45RA− CCR7+/ CD4+ CD62L+ CD45RA	CD197+									antigen-experienced central memory	[11][20][39]
								CM Th1															
								CM Th2															
								CM Th3															
							Effector memory help			30.5 ± 13.3% [89]	CD3+/CD8−/CD4+/CCR7−/CD45RA−	CD45RA− CCR7−	CD197−			T4eff.mem	1			√		effector subtypes that exist in resting or a	[11][20][39]
								EM Th1															
								EM Th2															
								EM Th3															
							Terminally differentia			Mean 16.9% in PB	CCR7−	CD45RA+ CCR7−/ CD4+ CD45RO+ CD45	CD197−										[11][20][88]
						Regulatory T cell (				5~10% of CD4 T c	CD4+ CD25+ FoxP3+ CD127−	CD14−/CD56−CD3+/CD4+CD8−/CD25+	CD127−			CD4+ CD25+ T	1		√				[20][27]
							aTreg (activated Tre				CD45RO+	CD25++CD45RA−(FOXP3hi)				aTreg	1			√			[20][25]
							rTreg (resting Treg c				CD45RO−	CD25++CD45RA+ (FOXP3lo)				rTreg	1			√			[20][25]
							sTreg (secreting Tre				CD45RO+	CD45RO+Foxp3low											[20][25]
																Non−classical	1			√			
																CD4+, CCR5+ (	1	GSM3203408				[34]	
					CD8+ T cell (Cytotox					5~30% [1][39], 13-	CD3+CD4 - CD8+					CD8+ T cell, Cy	1	GSI	√		√	CD8 T cell - to be extremely heterogenou	[39][68]
						Naïve CD8+ T cell				37.6 ± 19.4% [89]	CD3+/CD8+/CD4−/CCR7+/CD45RA+	CD14−/CD56−CD3+/CD4−CD8+/CD45RA	CD197+			T8naive, CD8+	1		√	√			[11][20][66]
						Activated CD8+ T cell					CD3+/CD8+/CD4−/CD38+/HLADR+												[66]
						Effector CD8+ T cell					CD3+/CD8+/CD4−/CCR7−/CD45RA+												[66][85]
						Memory CD8+ T cell																	[84][88][89
							Central memory cyto			6.25 ± 4.40% [89]	CD3+/CD8+/CD4−/CCR7+/CD45RA−	CD14−/CD56−CD3+/CD4−CD8+/CD45RA	CD197+										[11][20][66]
							CD8 PMCD8 PM																
							Effector memory cyto			29.9 ± 12.3% [89]	CD3+/CD8+/CD4−/CCR7−/CD45RA−	CD14−/CD56−CD3+/CD4−CD8+/CD45RA	CD197−			T8eff.mem	1			√			[11][20][66]
							Terminally differentia			26.2 ± 18.1% [89]	CCR7−	CD14−/CD56−CD3+/CD4−CD8+/CD45RA	CD197−										[11][20][88]
						RegT (CD8+)	CD8+CD28 - Treg																
							CD8+ CD28−limited Treg																

Figure 18. The metadata for PBMC ontology building, based on selected PBMC SCT data.

In this ontology metadata, each cell type (subtype) has frequency, phenotype marker, function and properties, data source, additional information, references, etc. categories for lineage tracing and literature tracing. Related information and referenced literature have been stored in repository. The hierarchical relationship of each data set can be clearly located with the taxonomy dendrogram in Figure 6. It is significant to interpret single cell classification results with PBMC ontology and background metadata information.

3.2 Multi-Dimensional Single-Cell Ontology: PBMC as An Example

Domain knowledge (prior biological knowledge) is significant to data, model/algorithm, parameters in single cell classification process. It can help to interpret and address machine bias from the perspectives of inaccurate assumptions to data labels and flawed data sampling where data is over- or under-represented in machine learning training data set.

Currently existing cell ontologies are not suitable for single cell classification, with deeper resolution in SCT technology and new evolving concepts in cell type definition and determination. Traditionally, there are different classification criteria for cell types, such as cell morphology, molecular-cell function (surface receptors, cell secretions, etc.), but these criteria are not always connected. In addition, the cell classification ontology, standard, and naming of cell types are not consistent across different studies, to a certain extent. There is often a phenomenon of cell type recognition based on molecular markers discovered in certain research, or cell type determination standards that are chosen at purpose or for convenience.

The existing classification of immune cells does not have a systematic and comprehensive classification standard, which makes it difficult for us to understand cell types and classify them with ANN models. The current cell ontologies focus on describe cell types based on traditional methods. The determination of cell identity, cell type, cell state, and cell fate has entered the era of digital quantitative definition of each individual single cell. Single cell gene expression can be sensitively affected by factors of multiple dimensions: from cell properties, organism properties, types of tissue, experimental settings, and data analytics. The classification of single cells urgently needs a systematic and formally defined multi-dimensional ontology.

With the quantitative defined single cell gene expression profiles, in this section, a multi-dimensional single cell ontology has been systematically described, with taking PBMC cell properties specifically as an example, referring to the existing literature and collected 10x SCT data. That gives a hierarchical, common, and controlled vocabulary prototype for single cell ontology. The PBMC cell properties has been designed to be one layer upper based on existing data, and one layer of subclasses beneath the classes of the existing data.

The following has written the multi-dimensional single cell ontology proposed. This work has been organized into a paper manuscript under reviewing.

3.2.1 Abstract

We propose a multi-dimensional cell ontology for single cell study, with PBMC as a specific example. It has described over 163 dimensions to category and characterize single cells, based on prior knowledge in immunology and single cell study domain. The multiple dimensions include cell types and factors affecting single cell gene expression level. This ontology can be used as a reference model to support with single cell data analysis, such as single cell classification.

3.2.2 Introduction

Ontology is a formalized representation of the definition of a group of concepts, and the standardized description of their attribute relationships, in a certain field. Ontology represents and describes two questions of concepts in a field – "what are they" and "what are their relationships". Ontology helps to strengthen the certainty and clarification of the nature of research objects or facts. It is the basis for the understanding of research data and research questions [178, 179].

In single cell study field, it requires an ontology to annotate and category single cells with hierarchical structure of multiple dimensions.

At the level of single cell, the cell gene expression can be affected by diverse elements: an inherent expression related to cell type, and influence of tissue location, organism properties, experimental settings, data analytics.

For example, dendritic cells from tonsil has different expression to dendritic cells from peripheral blood [16, 65]; T cell gene expression can be changed by methanol fixation [16, 65]; the single-cell transcriptomics (SCT) technology platform (e.g. 10x Genomics v2, v3) has a greater impact on the similarity of cell gene expression than the cell type itself [59]; the gene expression profile of PBMC in chronic lymphoid leukemia (CLL) patients has changed significantly over time and treatment [29].

In domain, currently, there are ontologies, such as Cell Ontology (CL) (cellontology.org) (an ontology for cell types) [180], Gene Ontology (GO) [181, 182], that have been constructed and written in a set of standardized principles of OBO foundry [183]. However, CL focuses on general concepts of cell types from prokaryotes to mammals, it does not have available subclasses underneath the class "PBMC". Further, it is derived from the subjects of life science and cell biology, it has generally described cell types with the perspectives of cell origin, and cell function, etc.

The advance in SCT has brought a need in categorizing a single cell based on the concepts from diverse dimensions – not only from cell type, but also considering dimensions in tissue and organism, experimental processing and data processing. It requires a hierarchical vocabulary of multi-dimensions to categorize SCT profiles. It can support the repeatability and reliability in SCT analysis.

This ontology supplies a structured and controlled vocabulary for single cell study. It determines distinct hierarchical categories and relationships for individual single cells. The ontology can be used as a reference for single cell classification, that helps SCT data being classified according to precise dimensions and compartments [184]. It can guide machine learning model and statistical analysis to find differential expression patterns of SCT data on each specific dimension.

To meet the need of an ontology in single cell study, we produce a multi-dimensional ontology model, based on dimensions of cell properties, organism properties, tissue types, experimental settings, and data analytics. In cell properties, PBMC has been taken as example to describe. The biological knowledge of the ontology is from immunology [36, 185] and SCT research field. The ontology is built according to principles of being clear, concise, informative, and reliable.

3.2.3 Construction and content

3.2.3.1 SCT study dimensions

Efficient SCT data integration and classification requires the ontology in multiple SCT study dimensions.

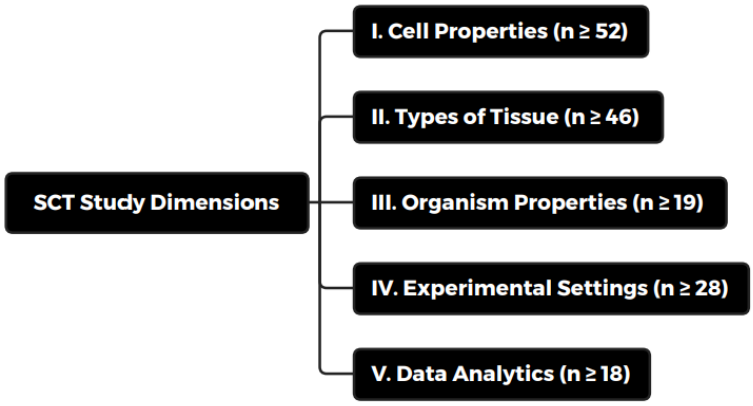


Figure 19. Five angles of SCT study multi-dimensions. The number in the figure shows the number of dimensions in each main angle. The ontology has over 163 dimensions in total.

Comprehensively, the SCT study dimensions include five main angles: cell properties, types of tissue, organism properties, experimental settings, and data analytics. These five main angles are the primary factors that need to be considered for SCT data integration, analysis, and classification. Each sub dimension in these five main angles can affect the specific gene expression level in individual SCT profile.

3.2.3.2 Cell properties and PBMC ontology

- **Cell properties**

First, specifically, in 'Cell Properties' angle, it has 12 sub dimensions, the first layer of 'Cell Properties' is comprised of 'Genetic lineage', 'Maturation status', 'Activation status', and 'Effector/memory' dimensions. 'Genetic lineage' is the dimension to decide SCT cell type in the view of cell lineage development. Based on our previous PBMC SCT classification study [65], it has two sub dimensions: 'non-PBMC' and 'PBMC'. 'PBMC' dimension has been structured in detail in Figures 21-26.

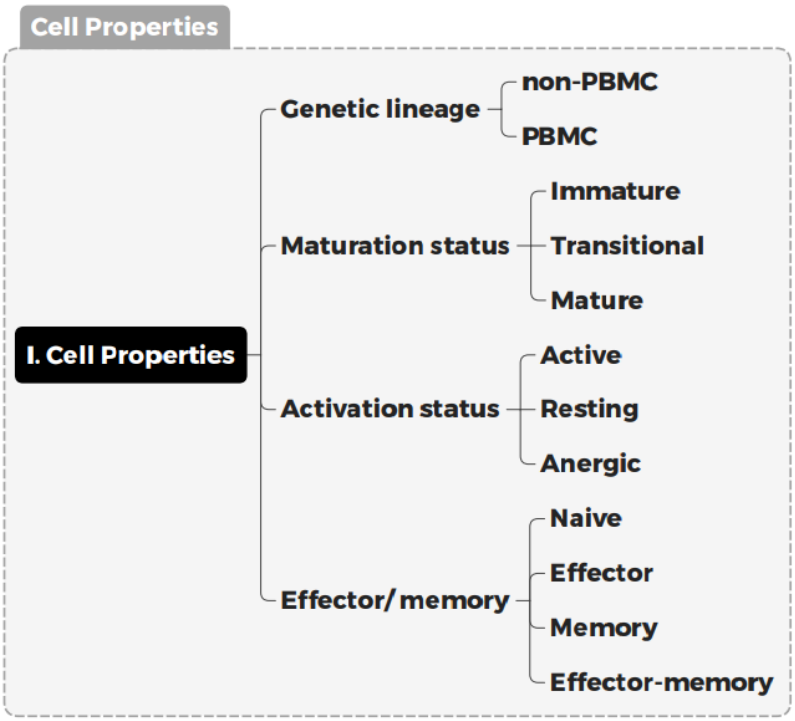


Figure 20. Dimensions in 'Cell Properties' angle. It includes subdimensions from 'Genetic lineage', 'Maturation status', 'Activation status', and 'Effector/memory', four dimensions.

Our ontology has set "the status of immune cells" as dimensions independent of "cell genetic lineage" (the dimension traditionally used to define cell types).

There are different views on the division of the hierarchy between "cell type" and "cell status" [11, 97, 184], and there are studies use "cell status" as a part of content in cell type determination and definition [186]. From the perspective of single-cell research and big data analysis, we have split the "cell lineage type" (named as 'Genetic lineage' in ontology) and "cell status type" as different dimensions to jointly define a gene expression profile of a specific cell population.

"Cell status" is an emerging concept for cell type classification [97]. The joint definition of cell type through "Cell status" and "Genetic lineage" is the development and continuation of the epigenetic landscape theory described by Waddington [187]. In our ontology, the branches of cell-fate decision points are jointly defined by multiple dimensions.

The characterization and determination of cell state is one of the key challenges in SCT [22].

In our ontology, 'Maturation status' has described dimensions in the maturation process, from immature, transitional, to mature. Immune cells gain mature status in specific immune organs, but

it has found their existing in periphery, during cell trafficking [185, 188].

The ‘Activation status’ dimension has divided immune cell into ‘Active’, ‘Resting’, ‘Anergic’, three compartments.

The ‘Effector/memory’ dimension is decided based on the time phase: whether the cells were stimulated by antigens, and the different differentiation stages they were in after receiving the activation stimulus. The ‘Naïve’ compartment refers to mature cells not exposed to antigen, ‘Effector’ refers to immune cells performing effector function with short life span, ‘Memory’ refers to cells performing similar phenotype to ‘Effector’ cells, while with long life span (up to several years).

- **PBMC ontology**

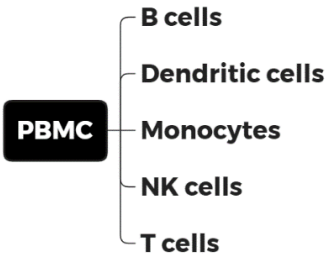


Figure 21. Five classes under the ‘PBMC’ dimension.

The dimension ‘PBMC’ consists of ‘B cells’, ‘Dendritic cells’, ‘Monocytes’, ‘NK cells’, and ‘T cells’, based on immunology prior knowledge [36, 185].

- **B cells**

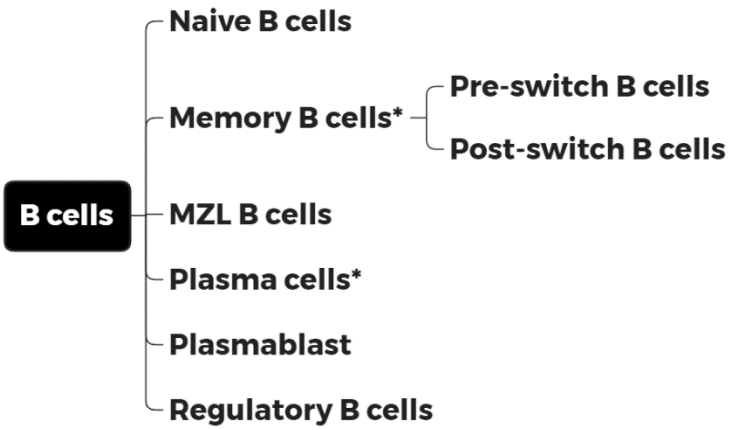


Figure 22. B cell ontology defined. ('MZL B cells' is the abbreviation for 'Marginal zone-like B cells'.)

In B cell dimension, the ontology has set six compartments - 'Naïve B cells', 'Memory B cells', 'MZL B cells', 'Plasma cells', 'Plasmablast', and 'Regulatory B cells' [188-192]. Immature B cells and Transitional B cells before complete maturation, are not described in the ontology.

After the maturation, naïve B cells enter the peripheral blood, they can be activated, effected, or brought to memory status, by self-antigens or hetero-antigens. Plasma cells are effector B cells, it is distinguished into two divisions based on different life span (short-lived; long-lived - from few months to lifetime) [193].

Pre-switched B cells and post-switched B cells (lgG+, lgA+, lgE+ memory B) are listed as two compartments of the dimension 'Memory B cells' [194].

Regulatory B cells perform the function of regulation in peripheral blood, it is proposed that any B cell has the capacity to differ into a regulatory B cell in human [195].

Other B cell groups with trace amount of cell numbers in blood are not involved in the ontology, such as B-1 cells (mainly in fetus blood), early plasmablasts, transitional plasma cells, etc.

While defining PBMC cell classes, we have found that PBMC cell types are largely defined by the types of specific cell surface markers (*e.g.* surface protein receptors, cluster of differentiation - CD markers), or, cells' secretions (*e.g.* immunoglobulin (Ig), cytokines, chemokines, granzymes, etc.). Examples can be found in DC-CL (a dendritic cell ontology) [196] and hemo-CL (a hemopoiesis cell ontology) [197]. This ontology has made effort to focus on the essential classes of cell types.

- **Dendritic cells**

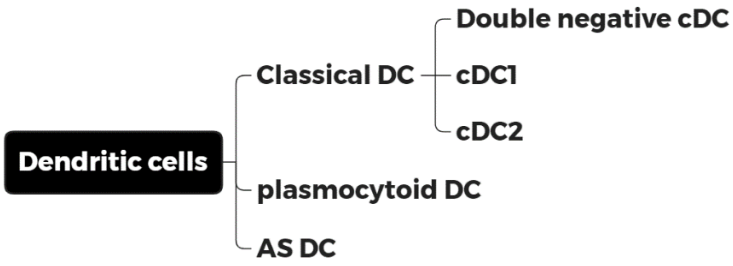


Figure 23. Dendritic cell ontology defined. ('AS DC' is the abbreviation for 'AXL+SIGLEC6+ DC cells'.)

The construction of dendritic cell dimension is based on prior knowledge [100, 198] and newly derived knowledge with SCT studies [48, 199]. In the ontology, 'Classical DC' shares the synonyms with "conventional DC", "myeloid DC".

The 'Classical DC' has the positive expression of CD11C. There are three subclasses under its dimension: CD11C+CD141+ DC (cDC1), CD11C+CD1c+ DC, and CD11C+CD141-CD1c- DC [48].

The 'plasmacytoid DC' positively expresses CD303 and CD123 marker [48]. The cDC can stimulate CD4+ T and CD8+ T in antigen-specific manner. The pDC produce type-1 IFN (interferon) as response to viruses [199].

The 'AXL+SIGLEC6+ DC' (AS DC) are newly defined in a DC SCT study [48], AS DC is unique to cDC or pDC. AS DC is isolated by co-expression of specific markers, such as, AXL, SIGLEC1/6, and CD22/SIGLEC2.

- **Monocytes**

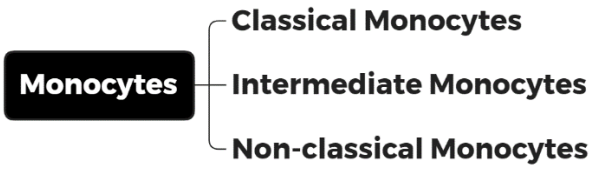


Figure 24. Monocyte ontology defined.

The monocyte dimension has three compartments: 'Classical monocytes' - CD14++CD16–, 'Intermediate monocytes' - CD14++CD16+, 'Non-classical monocytes' - CD14+CD16++ [48, 100]. The newly defined "Mono3" and "Mono 4" subtypes [48] are not listed in the ontology, given the consideration of further verification on reproducibility.

- **NK cells**

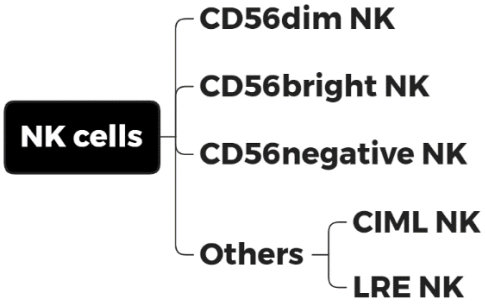


Figure 25. NK cell ontology defined. ('CIML NK', 'LRE NK' are the abbreviations for 'cytokine-induced memory-like NK cells', and 'population with low ribosomal expression NK cells', respectively.)

In NK cell dimension, there are four subclasses: 'CD56dim NK' - CD56+, 'CD56bright NK' - CD56++, 'CD56negative NK' - CD56–, and 'Others' [186, 198, 200, 201].

CD56bright NK and CD56dim NK both have two divisions: CD16– and CD16+. CD56brightCD16-, CD56brightCD16+, CD56dimCD16+, are, regulatory NK, intermediate NK,

effector NK, respectively.

Inside of CD56dimCD16+ compartment, there are two further partitions: CD56dimCD16+CD57– and CD56dimCD16+CD57+ [186]. The CD56dimCD16+CD57+ NK cells are terminally differentiated mature NK cells, with high cytotoxicity. Its reference range is around 12.2% of the total NK cells [186].

The ‘CD56negative NK’ is also termed as “inflamed NK” or “Type-1 IFN responding NK”. It is closely related to CD56dim cells while it has diminished cytolytic capacity [186, 200].

In the compartment of ‘Others’, ‘CIML NK’ and ‘LRE NK’ have been listed. The ‘CIML NK’ is strongly activated NK cells, it is similar to CD56dimCD94high intermediary NK cells, it is a hybrid between CD56dim and CD56bright NK cells [186, 200]. The ‘LRE NK’ is resembling to CD56dimCD16+CD57+ NK cells, while it has significantly reduced ribosomal expression. It is reminiscent of cells undergoing senescence or quiescence (termed as “ribophagy”) [186, 200].

There is a group of “adaptive NK cells” found in the NK SCT study [186], but not listed in the ontology, given the concern of reproducibility.

- **T cells**

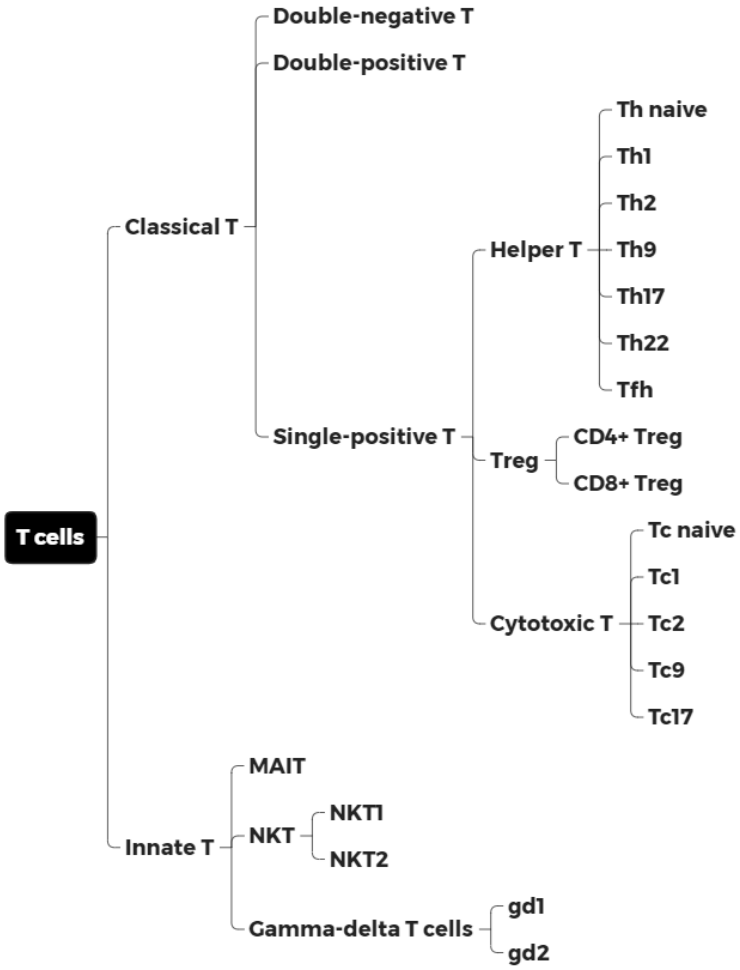


Figure 26. T cell ontology defined. ('Tfh', 'Treg', 'MAIT' and 'NKT' are the abbreviations for 'T follicular helper cells', 'regulatory T cells', 'Mucosal associated invariant T cells' and 'Natural Killer T cells', separately.)

In T cell dimension, there are two main compartments: 'Classical T' and 'Innate T' [202].

The ontology has set 'Double-negative T' - CD4– CD8–, 'Double-positive T' - CD4+CD8+, and 'Single-positive T' - CD4+/CD8+, compartments under 'Classical T', based on T cell lineage commitment [202]. Progenitor T cells experience T-cell receptor (TCR) gene rearrangement, thymus positive selection (MHC I, II) and negative selection (self-tolerance) to obtain single positive expression [185].

Under ‘Single-positive T’, based on the type of expressed surface receptors and the function, it has been divided into ‘Helper T’ - CD4+, ‘Treg’ - CD4+/CD8+, and ‘Cytotoxic T’ - CD8+. The ‘Helper T’ has set subdivisions including ‘Th naive’, ‘Th1’, ‘Th2’, ‘Th9’, ‘Th17’, ‘Th22’, and ‘Tfh’ [203, 204]. The ‘Cytotoxic T’ has subdivisions as ‘Tc naïve’, ‘Tc1’, ‘Tc2’, ‘Tc9’, and ‘Tc17’ [203, 205].

Effector Th1, Th2, Th17 can secret cytokines and have functions in cellular/humoral immune response. Naïve CD8+ T cells can be activated by effector helper T cells into effector cytotoxic T cells (CTL) [185]. In few cases, CTL can also be the effector CD4+ T cells [206].

Treg cells highly express CD25 and the transcription factor Foxp3, it is also labeled as CD4+CD25+Treg [198]. In the adaptive immune response, it can perform negative regulation function (as opposed to Th cells), through direct contact or the secretion of cytokines. Treg cells can turn other cells from an active status to a resting status. The CD8+Treg and Treg of other phenotypes have also been found [207, 208].

The ‘Innate T’ compartment includes ‘MAIT’, ‘NKT’, and ‘Gamma-delta T cells’. The ‘NKT’ and ‘Gamma-delta T cells’ compartments have subdivisions – ‘NKT1’, ‘NKT2’, ‘gd1’, ‘gd2’, respectively [209]. The ‘NKT1’ is also referred to as “invariant NKT cells” (iNKT) [210].

The ‘Innate T’ compartment is part of innate immunity of human body, as well as the ‘Dendritic cells’, ‘Monocytes’, ‘NK cells’ compartments. The ‘B cells’ and ‘Classical T’ compartments have functions in adaptive immunity.

The similarity between compartments “T cells”, “NKT cells”, and “NK cells” can lead to 2~3% of misclassification of T cells and NK cells, based on SCT data and supervised machine learning model [65, 146].

3.2.3.3 Organism properties

The ‘Organism properties’ angle has described at least 19 dimensions that can affect SCT cell gene expression profile, from the perspective of organism.

The dimension ‘Individual Genetic Differences’ represents factors influencing SCT profiles in gene level, from genetic background (in nature), to environmental exposure (acquired), and others.

Reference intervals and gene expression level of immune cell subsets can be different by regions,

populations, and ancestries [211-214], these factors conclude into 'Genetic background'.

'Environmental factor exposure' mainly refers to individual differences influenced by epigenetic modifications, such as industrial chemicals, heavy metals, air pollutions, temperature, humidity, light, ultraviolet radiation, mutagens, pharmaceuticals, vaccine [215], dietary components, alcohol, smoking, stress, sleep deprivation, behaviors, lifestyle, etc. [216-218]. Exposed to different environmental conditions, can make phenotype polymorphisms in genetically identical organisms.

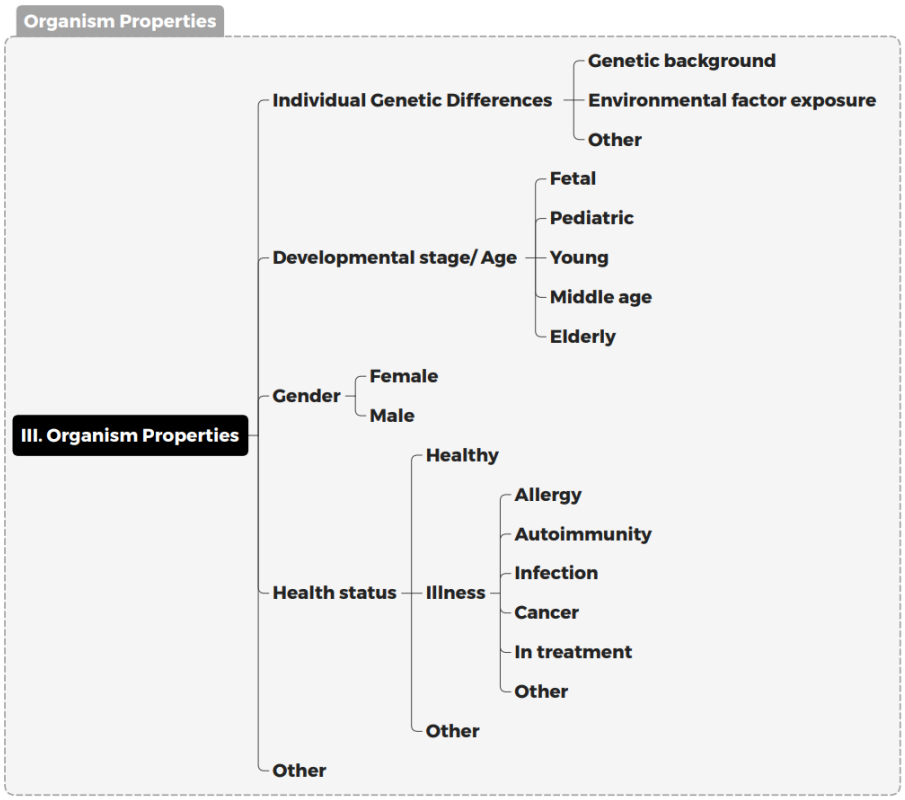


Figure 27. Dimensions in 'Organism Properties' angle. It includes subdimensions from individual differences, age, gender, to health status.

The influence of 'Developmental stage/Age' and 'Gender' on sample immune cell differences have been observed, as found in previous studies [163, 213, 214, 219, 220]. In the ontology, five compartments – 'Fetal', 'Pediatric', 'Young', 'Middle age', and 'Elderly', have been set under the dimension 'Developmental stage/Age'.

‘Healthy’ and ‘Illness’ dimensions can affect immune cell expression largely. The same type of cells can have specific gene expression in allergy [221, 222], autoimmunity [223, 224], infection [225, 226], cancer [227, 228], or, treatment [229], etc. conditions [163]. The change of PBMC gene expression in CLL patients with the process of treatment has been confirmed [29].

Other circumstances such as chronic disease [230], pregnancy [231] are also considered.

3.2.3.4 Types of tissue

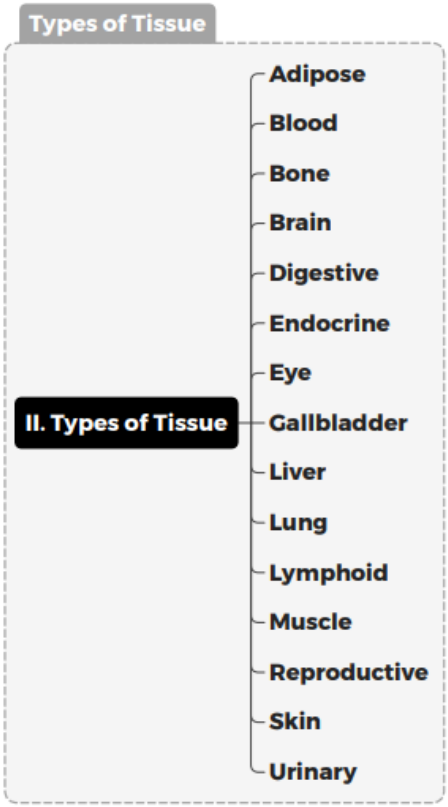


Figure 28. Division from the perspective of tissue type.

The settings of dimensions under “Types of Tissue” is done based on SCT data analysis practice and convenience, developed from views on traditional classification of anatomy, - the systems, organs, tissues, cells.

The construction of the dimensional hierarchy adopts the top-down principle.

The enumeration of dimensions based on different locations of organs and tissues conforms to the law of permutation and combination. The ontology only lists the partial types of tissues based on collected SCT data. The purpose of enumeration is to demonstrate a multi-dimensional model, rather than exhaustively list all types of organs and tissues.

Organs and tissues with available standardized SCT data include, “whole blood”, “PBMC”, “liver”, “lung”, “gallbladder”, “spleen”, “tonsils”, “breast”, “bone marrow”, “thymus”, “lymph nodes”, etc. In PBMC SCT data analysis, a common scenario is that less data comes from purified PBMC cell samples (such as only B cell samples or T cell samples), and more data are derived from PBMC mixtures or whole blood samples. This leads to the difficulty of PBMC cell splitting and the unavailability of the PBMC classification based on SCT data.

Another common situation is that, reading literature related to experimental data can find that many data samples marked as "peripheral blood" in the SCT database may come from tissues (such as "liver", "spleen", etc.), rather than the circulating blood on the periphery - in the traditional meaning. The definition of “peripheral blood” is related to the classification and analysis of PBMC. The SCT expression profiles of peripheral blood in different tissue environments are heterogeneous.

In particular, in PBMC SCT classification based on artificial neural networks (ANN), when adding tissue-residential dendritic cells (DC) data (from tonsil) to the training set [16], it can directly affect the accuracy of the classification model.

The studies [16, 65, 146] have shown the fact of SCT data vacancy on certain tissue type and the importance of clarifying specific sample tissue source in SCT analysis.

3.2.3.5 Experimental settings

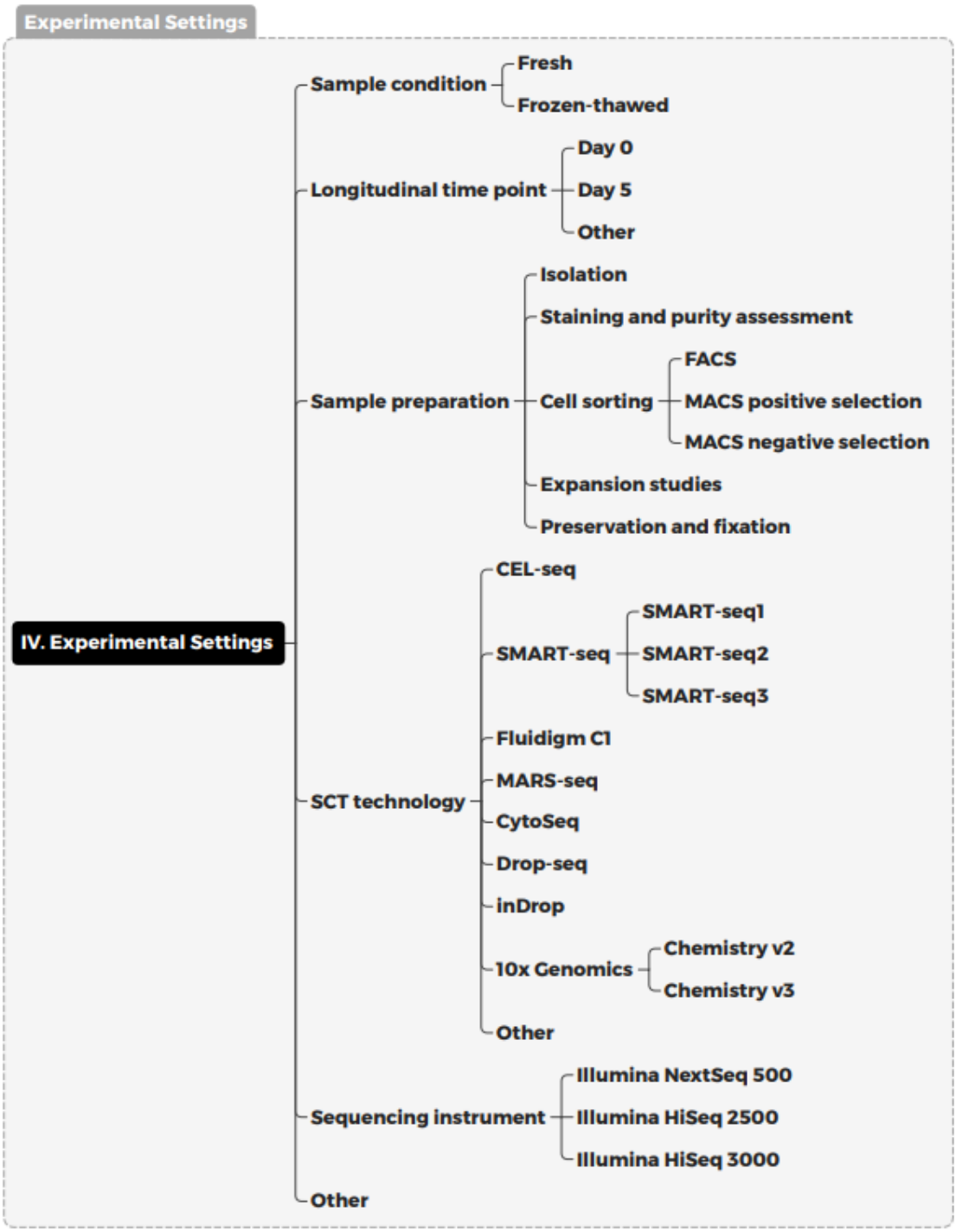


Figure 29. Dimensions of experimental settings involved in SCT data analysis.

Sample isolation, fixation, storage, sorting, processing steps in SCT experiment can affect the gene expression of measured cells [232].

A typical example is that, SCT data of T cells with methanol fixation [233] has apparently influenced the classification accuracy of ANN models [65].

It is very important to establish rigorous standard operating procedures (SOPs) and characterization methods for SCT data, that can avoid the introduction of technical variables in downstream analysis as much as possible. The data of the same cell type generated by different experimental procedures may not be comparable and reproducible.

- **Storage, temperature, and time**

SCT experimental material can be sampled with different conditions (e.g. fresh samples extracted from donor, or frozen-thawed samples received from sample library/biobanks).

Processing cell samples immediately after collection or within 24 hours [234] is the expected way to obtain satisfied gene expression data. An over high temperature can affect the vitality and functional activity of PBMC [235].

Due to the complexity of blood sample collection and the lack of samples, it is difficult to obtain fresh blood samples and process them in time. Low-temperature storage after collection has become one of the potentially acceptable solutions.

Transport temperature [236], storage temperature [50, 52, 237] and storage time can greatly affect the gene expression pattern of cells [232, 234]. Different storage temperatures can activate or inhibit the expression of certain genes [50].

Long-term low-temperature storage cannot prevent the degradation of RNA in frozen or refrigerated samples. Long-term low-temperature storage can lead to a decrease in cell viability and a decrease in the number of living cells [238], at the same time, the composition and function of cells can be changed [238].

At present, the preservation [238], thawing, and RNA extraction methods [239] of frozen blood samples are constantly being optimized.

- **Cell sorting**

Markedly, due to the advancement and particularity of single cell technology, the impact of different cell sorting techniques on PBMC gene expression also needs to be considered carefully.

The current mainstream cell sorting techniques include fluorescence-activated cell sorting (FACS), magnetic-activated cell sorting (MACS) positive selection and negative selection.

Control and evaluation of the cell sorting process is very important to preserve biological characteristics (gene expression level, cell function and differentiation status) of sampled cells [240, 241]. The influencing factors usually come from the stimulation, perturbation, stress or injury to cells during cell sorting [242]. Stress response genes may be upregulated by FACS sorting devices. Compared with magnetic positive selection, the gene expression characteristics between cells separated by magnetic negative selection and FACS can be more similar [243].

Expansion studies involved in functionally selected cells should be split from normal studies, in preparing data for SCT analysis [16, 65].

In the five-classification of PBMC, the gene differential expression coming from cell sorting method has been covered by differential expression coming from cell type. It has not significantly affected the model learning process and prediction performance [146]. Its impact on classification of sub cell types remains to be studied further.

- **Different SCT techniques and sequencing instruments**

Benchmark tests and evaluations [18] of different SCT protocols have shown they have different abilities to capture biological information in samples, reflecting on read structure and alignment, sensitivity, and range of multiple peaks (data distribution).

Currently the most widely used SCT technologies are 10x Genomics (10x) and Smart-seq2.

Smart-seq2 technology is a full-length sequencing, plate-based, low-throughput method, while 10x is a 3'-end or 5'-end sequencing, droplet-based, high-throughput method.

The Smart-seq2 protocol has advantage in higher sensitivity - it can detect a greater number of transcripts (larger exon read ratio, larger median value in distribution [18]), can detect more low-abundance rare transcripts, and RNA splicing isoforms [17]. Low-throughput methods are much

superior than high-throughput methods for research that demands the maximum sensitivity [244].

But it has a higher proportion of mitochondrial genes detection and a data combination that is more similar to bulk RNA sequencing.

In high-throughput methods, 10x has performed the best [18]. 10x can detect the most UMIs and genes in each cell, also can detect more long non-coding RNA (lncRNA) in a cell [18]. It can cover a huge number of cells and have demonstrated good performance in recognizing rare cell types [17].

However, 10x technology has 'dropout' phenomenon, it has higher background noise and random capture for low-expression RNA. The 'dropout' comes from the missing in capturing, reverse transcription, and sequencing.

Compared with 10x Genomics (v2), 10x Genomics (v3) has higher sensitivity in capturing RNA molecules. In terms of restoring the quantity of rare cell types, 10x Genomics (v2) has better capability than 10x Genomics (v3) [18].

For a same cell type, different technology platforms can produce SCT profiles with different data distribution and data structure characteristics [17, 18]. The technology platform can even affect the similarity of gene expression profiles more than the cell type itself [59].

Presently, supervised learning SCT cell classification has focused on data generated by 10x technology [146]. The SCT data generated by other technologies can to be collected and standardized, to further verify the generalization of the classification model.

The difference in sequencing instrument also has impact on the sequenced data [245]. Studies have shown that there exist differences in sequence deviation patterns within different sequencing platforms [246]. In contrast, the Illumina HiSeq series may have more significant preceding-base bias.

Standardization and quality control of experimental procedures are very important to produce usable and reproducible SCT data.

It is worth emphasizing that the sequencing depth and read length can have impact on SCT profiles [247]. For non-UMI-based SCT protocols, genes with short read length are more captured.

Adequate read length and sequencing depth can limit the technical noise [247]. However, too large sequencing depth can make the measured SCT profiles of different cells more similar.

SCT protocols based on UMI fragment reading (such as 10x Genomics) is not affected by read length.

3.2.3.6 Data analysis

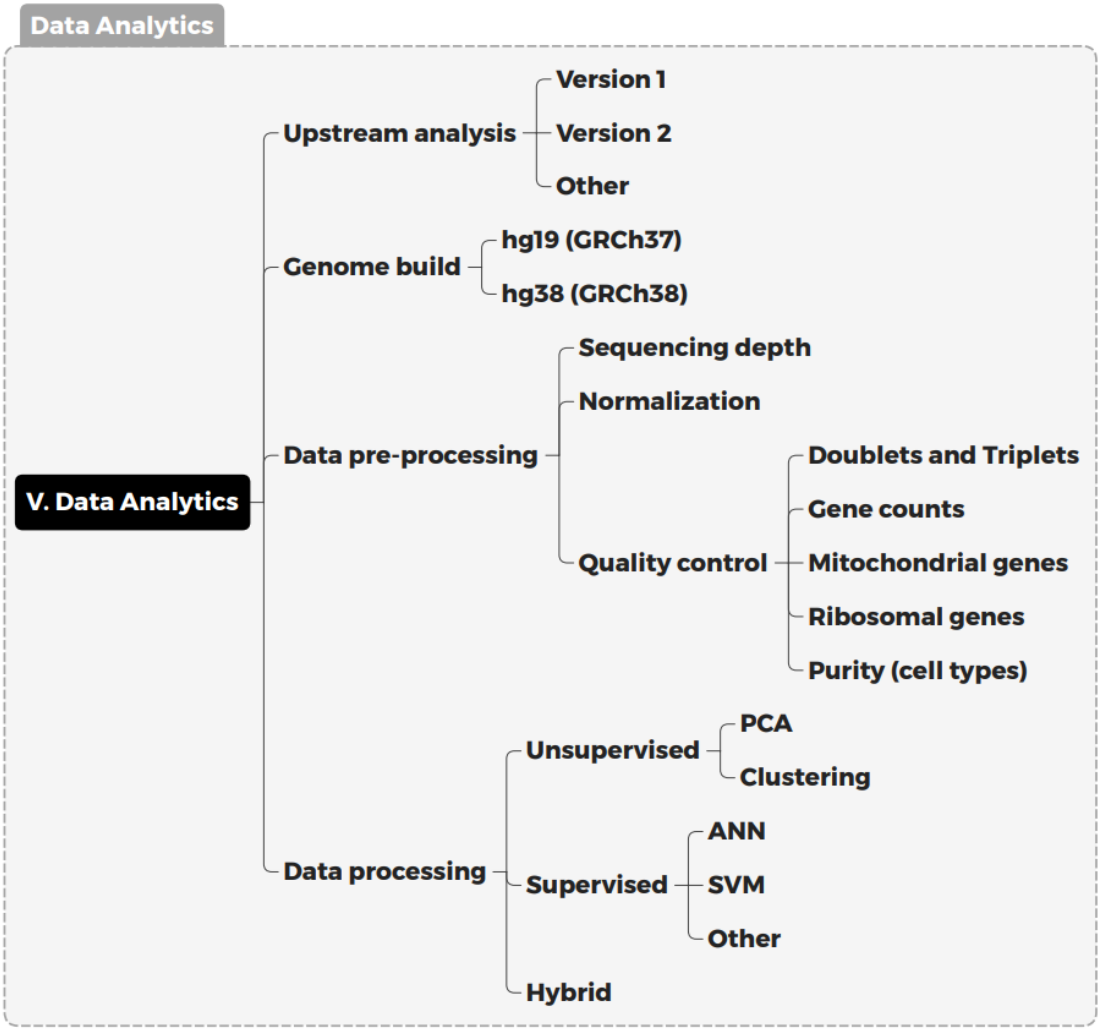


Figure 30. Dimensions in data analytics of the ontology.

Processing steps in data analysis creates data characteristics in more dimensions.

As for 10x Genomics protocol, upstream analysis to raw sequencing data can be performed with

software Cell Ranger, that has different versions from v1 to v6.

In the process of aligning the reads with the reference genome, there are different genome versions to choose from.

In the data pre-processing before downstream analysis, the parameters and the thresholds in the steps for normalization and quality control – on gene count number, mitochondrial genes, ribosomal genes, cell type purity, etc. can create various formatted results. That indicates new dimensions in SCT.

The 'cell type purity' here refers to the data-based, instead of purity assessment in cell sorting, one example is removing red blood cells (RBC), that recognized by unusual high expression of RBC genes, from PBMC SCT data.

Different clustering algorithms and annotation references in unsupervised data processing can produce distinct results in the numbers and categories of cell type.

For supervised classification methods, the training data quality and label reliability can decide the model behavior.

The downstream cell classification that minimizes the deviation from the real fact requires a strict and standardized SCT data process, including all the dimensions both in the Experimental Settings and in the Data Analytics.

3.2.4 Utility, conclusion, and discussion

This ontology uses controlled, structured vocabulary to summarize the general categories and multiple dimensions in SCT data analysis, with PBMC cell subtypes as an example.

It mainly describes three parts: the first is the name and determination of the cell type, the second is the multi-dimensional identity of each cell type, and the third is the SCT identification marker (protein marker and RNA marker) of each cell type.

This ontology represents a multi-dimensional model for SCT study and demonstrates as a reference for PBMC single cell classification. It has described five main angles in the ontology. The dimensions described are the basic perspectives of SCT gene expression characterization, they should be considered carefully before conducting data analysis.

SCT data downstream analyses (in particular, cell classification, cell heterogeneity analysis, etc.)

involve the discrimination of general categories and dimensions of single cells. Previously, the type of cell is commonly defined by morphology, function, and type of surface receptors. The resolution of single cell requires a multi-dimensional definition of the cell type. In practice, it can be found that the type or identity of a cell is usually determined by the intersection of different dimensions, that is a very common situation. Changes in one dimension can synergistically introduce switch in another dimension.

The ontology has been built based on fact and logic. A clear and explicit SCT ontology can help accelerate the construction of SCT analysis automation [248] and scale down the misclassification in SCT cell classification [65].

The ontology needs to be continuously updated and maintained. The current multi-dimensional model is mainly constructed based on domain prior knowledge and practical experience in analysis. The ontology also requires further suggestion come from experts in the field. Other new dimensions, such as new knowledge derived from SCT analyzed data, need to be continuously added to the ontology.

The ontology paradigm represented in this study can also be used in other genomics, proteomics, metabolomics research fields.

3.3 Classifier and Performance Assessment Methods

3.3.1 Classifier - ANN

A fully connected feed-forward artificial neural network (ANN) has been deployed for the study. The ANN system used in this study is illustrated in Figure 31.

The multi-layer perceptron classifier MLPClassifier of scikit-learn [249] python library (functions from the class "sklearn.neural_network.MLPClassifier", available at www.scikit-learn.org) has been used for software implementation.

The ANN architecture consists of one input layer, one hidden layer and one output layer (Figure 31 B). The input layer has 30,698 input units corresponding to the 30,698 genes in our standardized SCT data sets (the rows in the sparse matrices).

The ten hidden nodes have been chosen to use after exploratory analysis that showed the best balance between the classification accuracy and training speed. The preliminary experiments have been accomplished with ANN architectures comprising 100, 50, 25, 10, 5, 2, and 1 hidden layer nodes [16]. It has been concluded that ten hidden nodes provide the best balance between the ANN model classification accuracy and the speed of training process. For example, for Cycle 1 data (in the study of the proof of concept, Chapter 6) the accuracy of cross validation of architectures with 1, 2, 5, and 10 hidden layer nodes have been 73.4%, 92.2%, 99.79%, and 99.85% respectively. Further increases of the number of hidden layer nodes did not improve prediction accuracy.

The output layer is composed of five output units (BC, TC, NK, MC, and DC classes) referring to the respective five PBMC cell types (B cells, T cells, NK cells, monocytes, and dendritic cells).

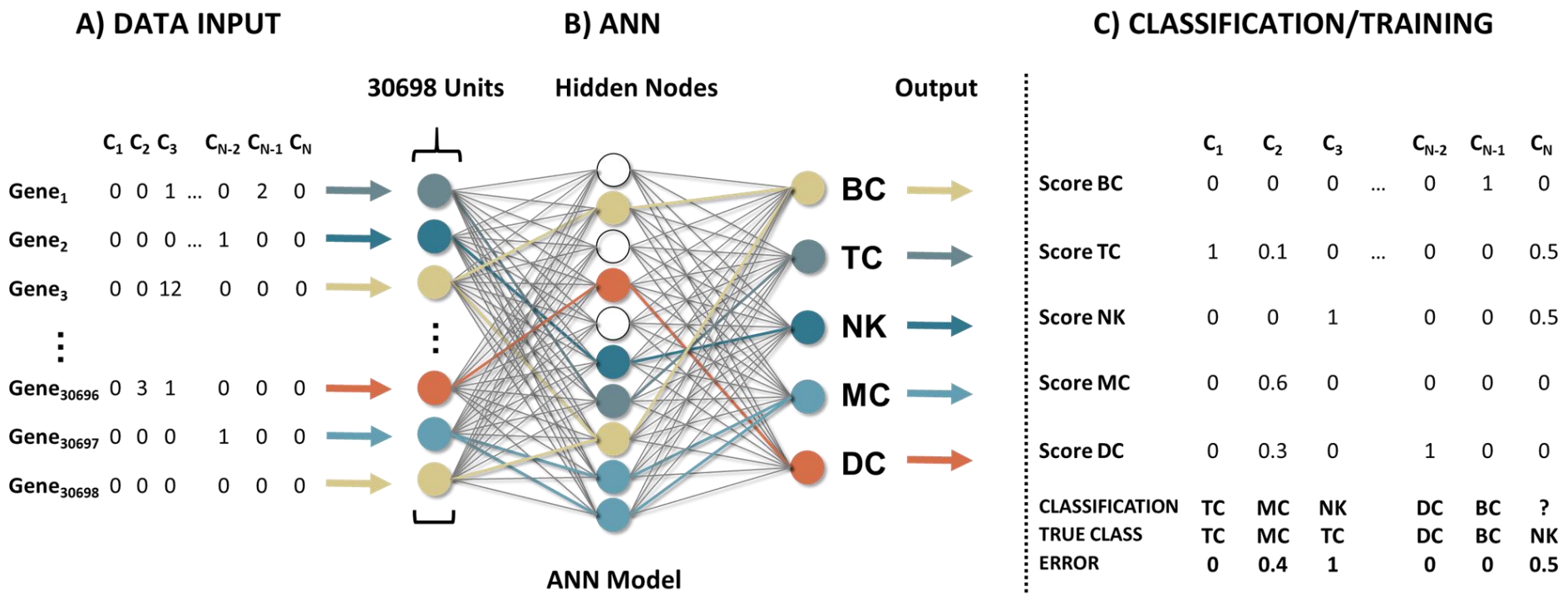


Figure 31. The ANN classification model architecture. The input data (A), ANN architecture (B), and the output data (C) are shown in this figure. The input data are in the form of sparse matrices where counts are represented by zeroes or positive integers. The architecture is fully connected ANN with 30,698 input units, 10 hidden layer units, and 5 output units, where output units correspond to classes representing major PBMC cell classes. The activation function ReLU has been used in this model, other parameters in detail have been documented in text below. The outputs are represented as matrices of output values that are used in training (by calculating errors) or for prediction of the class of cells of unknown type.

The activation function of the hidden layer nodes is rectified by linear unit ReLU, $f(x) = \max(0, x)$. The training data splitting minibatches of the size 200 is used to train the ANN model. The Adam algorithm [250] is used for first-order gradient-based optimization to train the neural network. The ANN model was set to random seed 42. The initial learning rate in the architecture is adjusted to 0.001 (10^{-3}).

The early stopping method has been performed for the prevention of data overfitting. In each ANN training process, 10% of the training data is put aside for validation while the remaining 90% of the data is used for ANN model training. The reaching point of ANN training stopping condition is set as when the prediction accuracy of the model on validation data sets is not improved for over ten continuous iterations (*i.e.* when the classification accuracy assessed by validation failed to improve for 11 iterations).

The training data is in the form of large matrices (N × 30,698), where N is the total number of columns – cells in each training step. Gene expression counts of 30,698 genes (Figure 31 A) are in the rows. The output consists of five real numbers obtained from each of the output units, and their sum is $V_{BC}+V_{TC}+V_{NK}+V_{MC}+V_{DC}=1$ (Figure 31 C). During training, the weights of the ANN are adjusted and after each adjustment the error is calculated as the sum of the absolute values of the difference between the expected value (one for the correct class, and zeroes for incorrect classes) and the actual score of the output units. The ANN training algorithm adjusts the weights between the nodes to minimize the overall output error. For classification, the true class of each cell is unknown, and the predicted class is determined by the maximum value of the five outputs (Figure 31 C).

The model has been trained with standardized SCT training sets, while tested with well-annotated high-quality testing sets. The model has recognized different transcriptional expression patterns across different cell types, that is learnt from training with well-labeled PBMC SCT data sets.

3.3.2 Assessment of classification performance

Certain assessment metrics have been used to evaluate and validate the performance of the model on PBMC classification. These are used to certify the understanding of the predictors' behavior and performance crosswise different training and testing steps.

3.3.2.1 Confusion matrix

A five-class multi-dimensional confusion matrix has been used for analysis of classifier performance, to present a complete picture of classification performance for all individual cell subtypes.

Confusion matrix records and reappears the classifier's prediction performance to each individual single cell in each experiment step. Confusion matrix is a two-dimensional digital matrix in which the row values on behalf of the cell number of each true class label, while the column values represents the cell number of prediction results voted and assigned by the ANN model (as shown in Figure 32). Confusion matrix can detect the trend of ANN classification performance, *i.e.*, it can identify if the trained model is frequently mislabeling one class as another. The classification result of each training and testing experiment step has been recorded in each confusion matrix for following analysis.

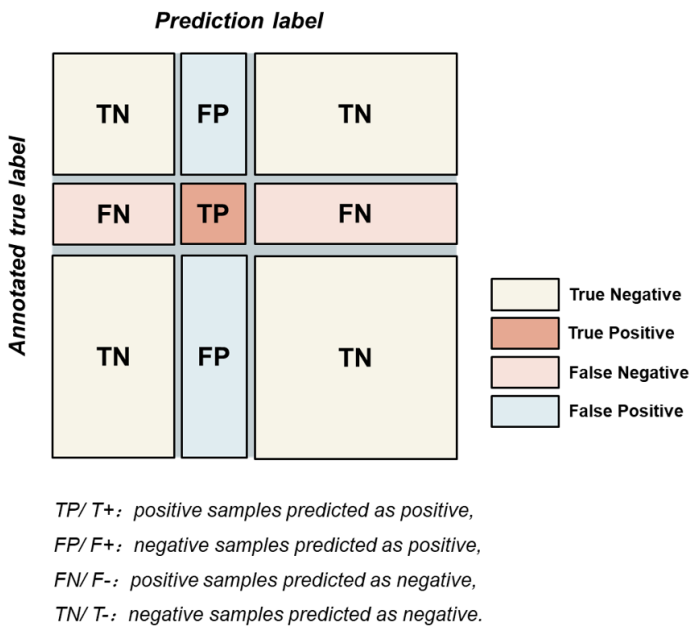


Figure 32. Illustrator of a confusion matrix. Confusion matrix is a visual model evaluation method, that consists of four situations to the result – true negative, true positive, false negative, and false positive. Metrics (Recall, sensitivity, specificity, precision, F1 score and overall accuracy) used to measure the capability of ANN classifier are sourced from confusion matrix. The detailed formulas and the relationship among these metrics have been explained as followed.

3.3.2.2 Appraisal indicators for comprehensive interpretation

The assessment metrics sensitivity (SE), specificity (SP), precision (PR) and recall (RE) as well as the harmonic mean, the F1 score have been measured in each confusion matrix to evaluate the classification performance of each cell class in each step. The formula of Sensitivity/Specificity (Formula 1), Precision/Recall (Formula 2), F1 measure (Formula 3), and the overall Accuracy (Formula 4), are following:

$$SE = \frac{TP}{TP + FN} \qquad SP = \frac{TN}{TN + FP} \quad (1)$$

$$PR = \frac{TP}{TP + FP} \qquad RE = \frac{TP}{TP + FN} \quad (2)$$

$$\mathrm{F1} = 2 \times \frac{PR \times RE}{PR + RE} \quad (3)$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

where,

TP – the number of true positives (experimental positives that are predicted as positives),

TN – the number of true negatives (experimental negatives that are predicted as negatives),

FN – the number of false negatives (experimental positives that are predicted as negatives),

FP – the number of false positives (experimental negatives that are predicted as positives).

The PR refers to the prediction result. It means the probability of true positive sample among all the samples predicted to be positive. PR can be confused with accuracy value, but they are two different concepts. PR represents the accuracy of the prediction to positive sample results, while the accuracy rate represents the overall prediction accuracy, including both positive samples and

negative samples.

The RE refers to the original sample. Its meaning is the probability of being predicted as positive in truly positive samples. PR and RE are a measure of the trade-offs. It is necessary to combine the results of the two indicators to find a balance point to maximize the comprehensive performance of classification.

The SE/SP values and the PR/RE values have been measured for each cell subclass as set in binary classifier, e.g. for B cells performance these values were measured for the result of B cells and non-B cells (union of DC, monocytes, NK cells and T cells). For the evaluation of incremental learning experiment design, the SE and SP value for each cell class in each periodic cycle were calculated to show the behavior of ANN classifier on each cell type during the procedure.

The SE and RE represent the same entity. Because it has performed multi-class classification, accuracy measure has been used for the assessment of overall performance, while F1 values are used for the assessment of performance in the classification of individual cell types.

The overall predictor performance has been assessed with the metric Accuracy (ACC).

The accuracy rate is defined as the percentage of the correctly predicted results in the number of the total sample (Formula 4). The accuracy value of each training and testing step has been calculated and recorded to validate the model classification performance on testing data sets.

In the result analysis procedure of the study – incremental learning (Chapter 6), the prediction result of dendritic cells had been put together into the prediction result of monocytes. The curve of ACC to testing data set classification results in different cycles (steps) can demonstrate the performance properties, robustness, and generalization of ANN model during incremental learning process (Chapter 6, 7).

CHAPTER 4 STUDY I – PROOF OF CONCEPT

This study has demonstrated the proof of concept of single cell classification done with supervised machine learning method ANNs and standardized SCT data of five cell types from PBMC samples. The work has been organized and published on the 2019 International Conference on Bioinformatics and Biomedicine (BIBM) [16]. This work was performed jointly with team colleagues. The metadata organization and training and testing sets preparation was performed by the author, the model setup was performed by the team colleague.

4.1 Abstract

The 27 human single cell transcriptomics (SCT) data sets have been used to develop an artificial neural network (ANN) model for classification of Peripheral Blood Mononuclear Cells (PBMC). We demonstrated that highly accurate models for classification of PBMC subtypes can be developed by combining multiple independent data sets to form training data sets. A significant data preparation effort was needed for building predictive models. Using a data set of ~120,000 single cell instances we showed the accuracy of classification of PBMC call of ~ 90%. Optimization techniques and addition of new high-quality data sets for model training are expected to improve PBMC subtype classification accuracy.

4.2 Introduction

This work has been demonstrated as the proof of concept that single cell classification can be done with purely supervised ML method ANN and standardized multi-source SCT data.

We standardized a selection of datasets that represent SCT profiles of major subsets of PBMC and trained artificial neural network (ANN) to classify five main types of PBMC cell subtypes. Given the rapid expansion of experimental data, the set of models generated in this study should be able to accommodate future, currently unknown cell types. Several research questions were pursued in this study:

Can we train an ANN on a set of data extracted from unrelated SCT studies and accurately classify PBMC cell subtypes?

How many different data sets are needed for developing accurate classification models?

Is it possible to generate accurate prediction models without feature selection or dimensionality reduction?

Is it possible to use tissue-resident immune cell subsets to accurately predict PBMC Cell subtypes of the same kind?

4.3 Materials and Methods

4.3.1 Data

Data were extracted from three sources, together with the metadata describing the samples and experimental conditions. We have collected, cleaned, labelled, and standardized 27 SCT data sets from multiple single cell gene expression studies. The labels corresponded to the PBMC cell subtypes – B cells, DC, monocytes, NK cells, and T cells. Each data sets only contain cells labeled as one specific subtype of PBMC. The number of datasets from individual sources are shown in Table 3. Nine datasets were from the 10x company demonstration data (10xS data set) [10], 13 datasets were from the GEO database (GEOS data set) [251], and five datasets from the Broad Institute (BroadS data set). The 10x data sets represented raw transcript counts for $CD19^+$ B cells, $CD14^+$ monocytes, $CD56^+$ NK cells, four sets of $CD4^+$ T cells, and two sets of $CD8^+$ T cells. The GEO datasets were extracted from Sample IDs GSM3258348, GSM2773408, GSM2773409, GSM3375767, GSM3087629, GSM3209407, GSM3209408, GSM3430548, GSM3544603, and GSM3478792. The Broad Institute datasets (BroadS) were extracted from the single cell study SCP345. Most of the data were in the Raw Count format, except for GSM3544603 and SCP345 that were log-transformed. We transformed back these two data sets to the same scale as others by rounding to the nearest integer the result of antilog transformation: $y = 2^x - 1$, where x is the previously log-transformed value from the source data and y is the antilog-transformed value approximating raw transcript counts. Since we had only a limited DC data (142 cells) that were extracted from PBMC, we also included SCT data of DC extracted from tonsils and tumor ascites (GSM3162630 and GSM3162632).

The summary report of the data sets is shown in Table 3. The total number of cells we used in this study is 121,281; the breakdown of cell numbers by PBMC subtype is shown in Table 4.

Table 3. The number of data sets used in this study.

Cell Type	Number of datasets			
	10xS	GEOS	BroadS	Total
B cells	1	1	1	3
Dendritic cells	0	2	1	3
Monocytes	1	3	1	5
NK cells	1	1	1	3
T cells	6	6	1	13
Total	**9**	**13**	**5**	**27**

Table 4. Total number of cells available for this study.

Cell Type	Total number of cells			
	10xS	GEOS	BroadS	Total
B cells	10,085	1,760	1,751	13,596
Dendritic cells	0	4,352	142	4,494
Monocytes	2,612	2,519	1,668	6,799
NK cells	8,385	309	1,394	10,088
T cells	64,347	13,613	8,344	86,304
Total	**85,429**	**22,553**	**13,299**	**121,281**

All data sets were cleaned and standardized. The genes across these data sets were named using dictionaries from different genomic builds including Genome Reference Consortium Human Builds 37 and 38 (GRCh37 and GRCh38) and their various patch releases. We mapped these different versions of the genomic builds to GRCh38 patch release 12 (GRCh38.p12). To make data sets easily comparable, we preserved the genes that were common across all the genomic builds represented across our studied data sets. Each standardized data set contains 30,698 genes. The rows of the data matrix represent genes (features) and the columns represent cells with the expression values of all identified transcripts. There are 30,698 rows corresponding to each feature while the number of cells (columns) in each data set range from 142 to >12,000. The BroadS data contains only 21,814 features. We mapped the values of these features to the standardized data set (30,698 genes) and set the missing feature values to zero.

We divided the data sets into training and testing sets. The GEOS data was divided into GEOS1

training set (8 data sets) and the TE1 testing data set (5 data sets). The testing set TE1 comprises a combination of high-quality data sets data sets annotated experimentally. The testing data set TE2 comprises manually annotated data sets from BroadS. To avoid confusion of terminology between biology and statistics, we consider term “sample” as biological sample that is represented by one or more data sets. Individual cell profile is called “single cell instance” or “instance”.

4.3.2 Study design

The study design involves several cycles of training and testing designed to assess the effects of diversification of training data as well as generalization properties of the trained models. The specific train-test cycles were:

- Cycle 1: Train ANN using 10xS data + tonsil-resident DC data, test using 2-fold cross validation (internal cross-validation)
- Cycle 2: Train ANN using 10xS + GEOS data, test using 2-fold cross validation (internal cross-validation)
- Cycle 3: Train ANN using 10xS + GEOS + BroadS/TE2 (all 27 data sets) data, test using 2-fold cross validation (internal cross-validation)
- Cycle 4: Train ANN using 10xS data + tonsil-resident DC, test using GEOS data set (independent experimental test set)
- Cycle 5: Train ANN using 10xS + GEOS1 data, test using TE1 (independent experimental test set representing all studied cell subtypes)
- Cycle 6: Train ANN using 10xS + GEOS1 + BroadS/TE2 data, test using TE1 (independent experimental test set)
- Cycle 7: Train ANN using 10xS + GEOS data, test using BroadS/TE2 (independent expert-annotated test set)

Cell class in independent experimental data sets is determined by experimental measurement using fluorescence-activated cell sorting (FACS) instrument. The cells in expert-annotated data sets were labeled using unsupervised clustering and analysis of features. They annotated cells at the level of sub-subclasses (seven subclasses of T cells, 2 subclasses of both B cells and monocytes, and a

single subclass of both DC and NK cells).

We consider expert-annotated data sets to be of very high quality. The order of cycles was determined arbitrarily, starting from company demonstration data sets, and data sets from GEO database that had raw transcript counts. After low accuracy of classification was achieved in Cycle 4 an additional data set was extracted from GEO for assessment in cycle 5. The final addition was an expert-annotated BroadS data set that was alternatively used in Cycles 6 and 7 as described earlier.

4.4 Results

4.4.1 Training results

The artificial neural network with the smallest training set was trained using more than 42,000 instances - labelled cell data (Cycle 1), while the largest training set had more than 110,000 instances. The training took between 20 and 60 epochs (iterations) before terminating. A typical learning curve displaying the changes in log-loss and validation score with respect to number of epochs is shown in Figure 33, indicating smooth convergence. Typical learning showed convergence at 20-40 cycles and the training terminated after 10 cycles without an increase in Validation Score (Figure 33).

4.4.2 Internal cross-validation

Two-fold cross-validation was performed on progressively increasing data sets. The smallest set was 10xS set (Cycle 1 – 85,429 single cell instances), the middle set was 10xS+GEOS (Cycle 2 – 107,982 instances), and the largest set with all data was 10xS+GEOS+BroadS (Cycle 3 – 121,281 instances). The overall internal cross-validation results shoved very high accuracy. Cycle 1 had 99.8%, Cycle 2 had 99.3%, and Cycle 3 had 98.9% correctly classified instances. The overall Cycle 1 and 2 results (data not shown) were very similar to the Cycle 3 results (Table 5). In Cycle 3, 1.5% of B cells, 2.7% of DC, 2.7% of monocytes, 3.7% of NK cells, and 0.6% of T cells were misclassified. The highest misclassification rate was for NK cells (3.5% of experimental NK cells classified as T cells), DC (2% of experimental DC classified as monocytes), and monocytes (1.4%

of experimental monocytes were classified as DC). These results were corroborated by additional classification performance metrics shown in Table 6.

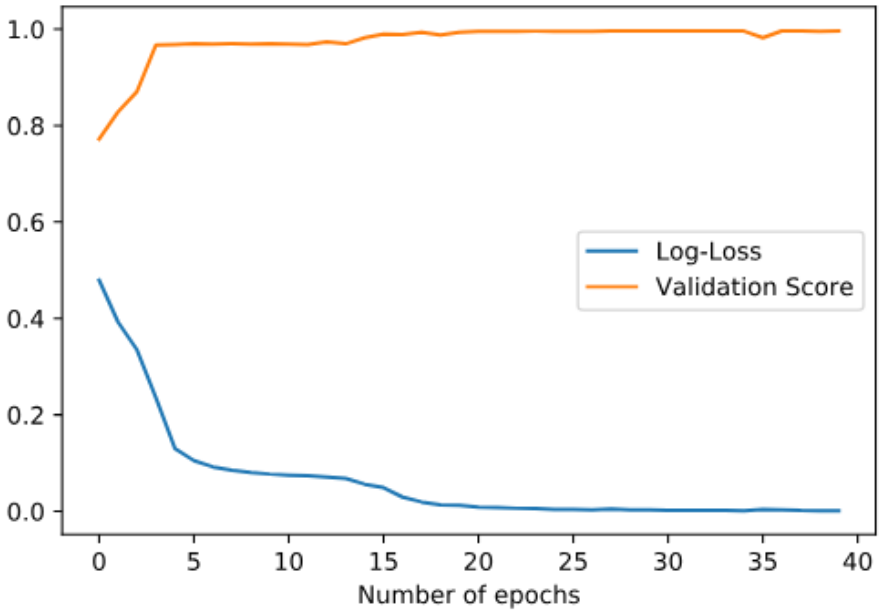


Figure 33. Representative ANN learning. The training stopped after 10 cycles of no improvement of validation score.

Table 5. Cycle 3 confusion matrix.

Predicted / Experimental	PBMC BC	PBMC+TO +TA DC	PBMC MC	PBMC NK	PBMC TC	SUM
PBMC BC	13,388	5	47	68	88	13,596
PBMC+TO +TA DC	1	4,374	88	1	30	4,494
PBMC MC	29	95	6,613	1	61	6,799
PBMC NK	9	3	4	9,719	353	10,088
PBMC TC	55	10	75	343	85,821	86,304
SUM	13,482	4,487	6,827	10,132	86,353	121,281

*BC: B cells; DC: dendritic cells; MC: monocytes: NK: NK cells; TC: T cells; TO: tonsil resident; TA: tumor-ascites resident; PBMC: peripheral blood mononuclear cells.

Table 6. Cycle 3 assessment metrics.

	PBMC BC	PBMC+TO +TA DC	PBMC MC	PBMC NK	PBMC TC
F1	0.990	0.976	0.974	0.961	0.994
PR	0.993	0.973	0.975	0.958	0.994
RE/SE	0.987	0.979	0.973	0.964	0.994
SP	0.998	0.999	0.998	0.997	0.986
ACC	0.989				

PR: precision; RE: recall; SE: sensitivity; SP: specificity, ACC: accuracy; F1: F1 score

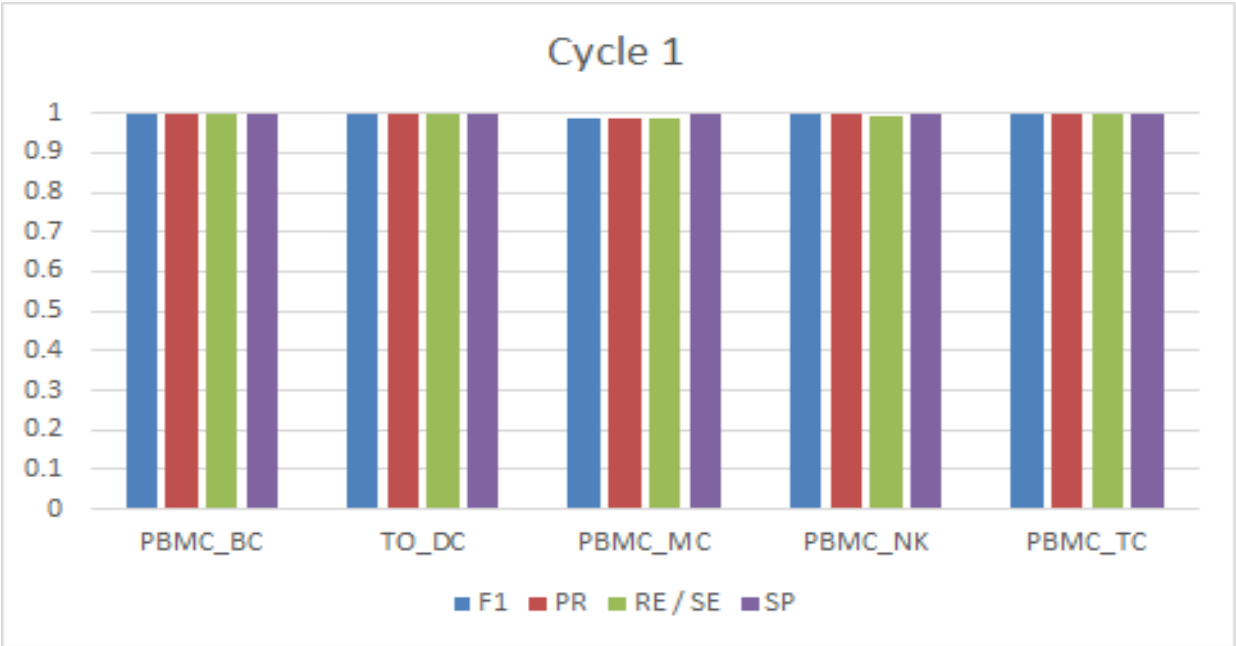
The cross-validation results indicate that the ANN learning is effective when we combine multiple data sets from different studies even if they are performed by different laboratories. If datasets are randomly split and a study is represented in both training and test sets, the misclassification rate for any cell subtype will be lower than 4%.

4.4.3 Prospective validation

After demonstrating that ANN can accurately classify cell subtypes represented in the training set (but not identical to the cell instances in the test set), we explored the generalization ability of trained ANN models. The process included diversification of training data by incremental addition of data sets.

In Cycle 4, we trained ANN using the 10xS + tonsil resident DC (TRDC) data and used the GEOS data set for testing. The GEOS data set did not contain TRDC data, but it contained tumor-ascites resident dendritic cells (TADC). This was done to explore whether PBMC resident DC can be predicted using DC from other tissues.

The same model that could perform highly accurate predictions using internal cross-validation (Cycle 1) could not predict previously unseen data sets with satisfactory accuracy. The accuracy of predictions in Cycle 4 was only 46.1% and none of the cell subtypes showed useful predictions (Figure 34).



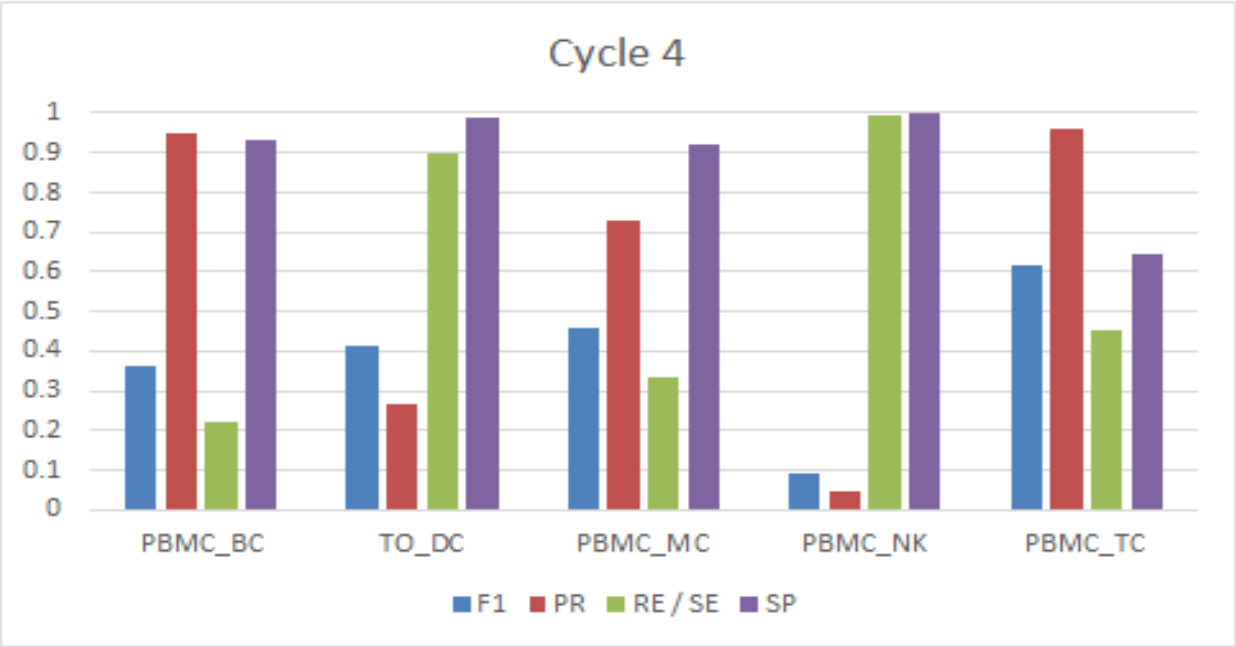
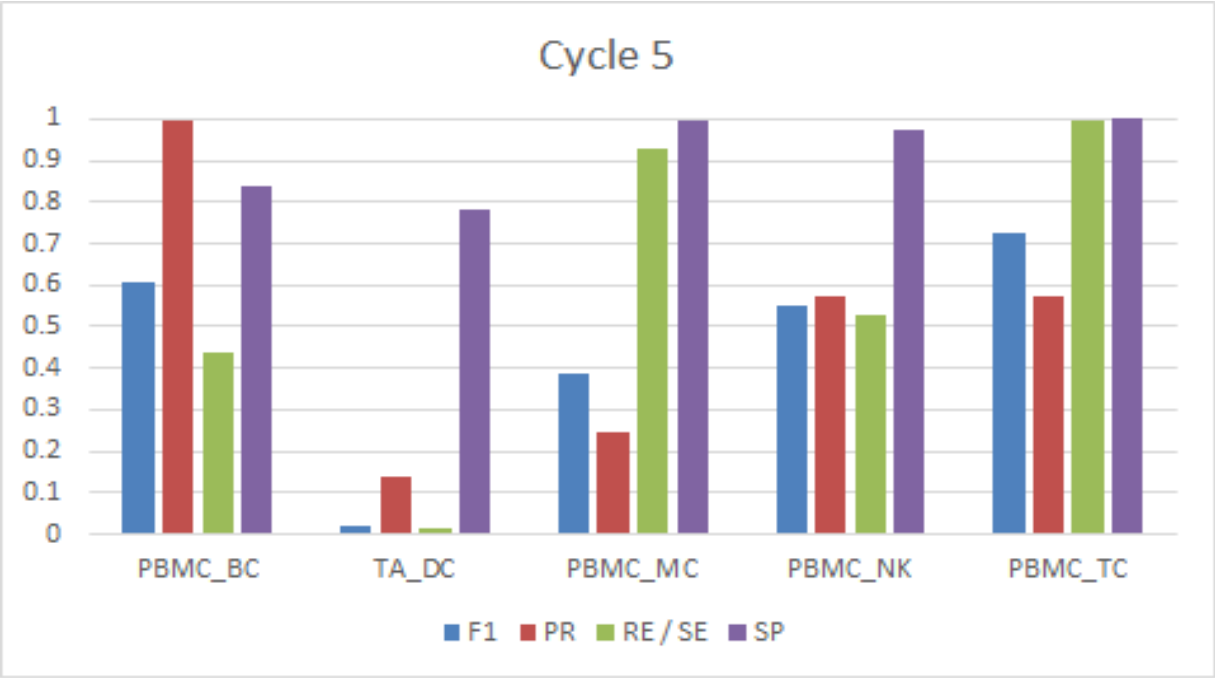


Figure 34. A comparison of classification performance for cycle 1 and cycle 4.

Cycle 5 involved splitting GEOS data (test set in Cycle 4) into GEOS1 data set and a smaller TE1 test set. GEOS1 was added to the 10xS to form a new training set, while TE1 was used to test

predictive performance in Cycle 5. In Cycle 6 we added BroadS data set to training set from Cycle 5 and tested using the same TE1 test set as in Cycle 5. The results show improvement in overall accuracy, 52.8% in Cycle 5 and 62% in Cycle 6. Although these were notable overall improvements (6.7 and 15.9% as compared to Cycle 4), the analysis of Cycle 5 data shows improvement of classification performance relative to Cycle 4 for T cells, B cells, and NK cells, whereas the performance declined for DC and monocytes (Figure 34 and Figure 35). The reason for this change was that majority of tumor-ascite resident DC were predicted as monocytes reducing accuracy of classification for both data sets. For Cycle 6, we added the BroadS data set to the training set from Cycle 5. The classification results for TE1 set show further improvement of predictive performance for B cells, NK cells, and T cells, whereas predictive performance for DC and monocytes remained low with the majority of tumor-ascite resident DC classified as monocytes (Figure 35).

The final step of this study involved training of ANN using combined 10xS + GEOS data set and testing using BroadS data set – Cycle 7. The advantage of this construction is that BroadS data set is derived from PBMC, including PBMC DC whose frequency is only 1-2% of the total PBMC. The result showed improvement of predictive accuracy relative to previous cycles, using a test set that is unseen by the trained ANN.



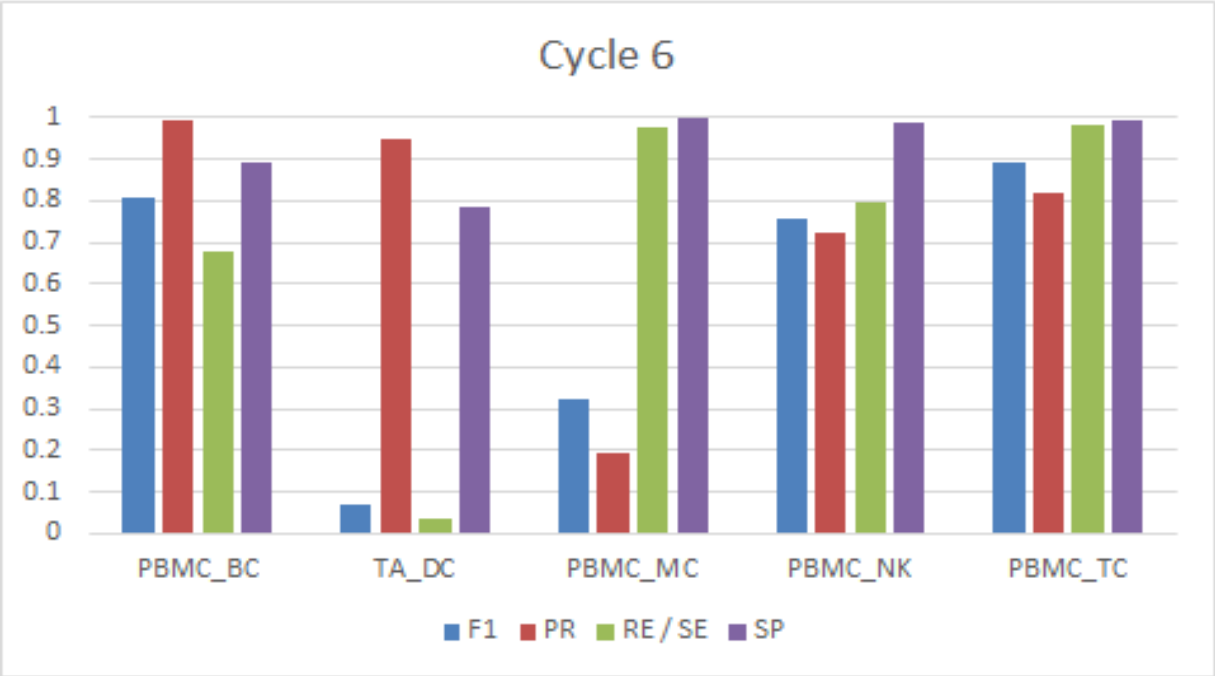


Figure 35. A comparison of classification performance for cycle 5 and cycle 6.

Table 7. Cycle 7 confusion matrix.

Experimental \ Predicted	PBMC BC	TA+TO DC	PBMC MC	PBMC NK	PBMC TC	SUM
PBMC BC	1,624	7	102	2	16	1,751
PBMC DC	0	69	72	0	1	142
PBMC MC	120	143	1,324	2	79	1,668
PBMC NK	23	11	4	1,110	246	1,394
PBMC TC	55	10	58	464	7,757	8,344
SUM	1,822	240	1,560	1,578	8,099	13,299

Table 8. Cycle 7 assessment metrics.

	PBMC BC	TA+TO DC	PBMC MC	PBMC NK	PBMC TC
F1	0.909	0.361	0.82	0.747	0.944
PR	0.891	0.288	0.849	0.703	0.958
RE/SE	0.927	0.486	0.794	0.796	0.93
SP	0.989	0.995	0.972	0.977	0.904
ACC	0.894				

The overall accuracy of Cycle 7 predictions is 89.4% (Table 7). In Cycle 7, 7.3% of B cells, 51.4% of DC, 20.6% of monocytes, 20.4% of NK cells, and 7.0% of T cells were misclassified. The highest misclassification rate was for DC (50.7% of experimental DC classified as monocytes), NK cells (17.5% of experimental NK cells classified as T cells), monocytes (8.6% of experimental monocytes were classified as DC and 7.2% of experimental monocytes classified as B cells), B cells (5.8% of experimental B cells classified as monocytes), and T cells (5.6% of experimental T cells classified as NK cells). These results were corroborated by additional classification performance metrics (Table 8).

4.5 Conclusions

We performed a cyclical refinement of ANN models by combining data from multiple unrelated studies into unified training set for prediction of PBMC cell subtypes. We achieved high overall accuracy of predictions 89.4%. We showed that ANN training using a limited number of related data sets, generated in the same study, does not generalize well and has low accuracy when tested with unrelated data sets. It is unclear how many diverse data sets are needed to achieve high accuracy of trained models. Our data indicate that two distinct B cell data sets (13,596 instances) produced an ANN model that performed well on an independent data set (F1=0.91, SE=0.93, SP=0.99). At the same time, two distinct NK data set (10,088 instances) produced an ANN model that had moderate performance on independent data set (F1=0.75, SE=0.80, SP=0.98). Having 10 or more data sets for each PBMC cell subtype appears to suffice for achieving a very high accuracy of trained ANN models, as seen for prediction of T cells (Table 8).

Furthermore, we have demonstrated that ANN models can be trained for high accuracy and excellent generalization properties without feature selection or dimensionality reduction. This will enable fine tuning of future training of ANN models to predict rare cell types without the need to redefine relevant features.

Our findings indicate that accurate prediction of PBMC-resident DC cannot be achieved by training using tissue-resident DC and tumor ascites DC. This finding indicates that SCT may be useful for developing diagnostic tests based on various tissue resident cell subpopulations, because each of them is likely to have own shared patterns of gene expression.

Finally, we noted that most of misclassifications involved bilateral misclassification of DC and monocytes and bilateral misclassification of NK cells and T cells. It is known that monocytes can differentiate into DC [252] making these two cell types a part of the same lineage. NK cells differentiate from the same precursor as T cells and B cells and may share molecular markers. At this point we cannot determine the reasons for high number of misclassifications of NK cells and T cells.

4.6 Discussion

To our knowledge, this is the first study that has applied supervised machine learning to data sets from multiple unrelated studies to classify cell subtypes. The training set in the final cycle exceeded 110,000 training instances.

We anticipate a rapid expansion of new studies that will share their data. This will create several challenges. First, there is a need for more systematic classification of cell subtypes [42] that will provide a new model of ontologies and cell taxonomies. Second, data sets are becoming larger and they appear with increasing frequency. We anticipate that GEO repository may have more than 100,000 data sets for 10x single cell transcriptomics as early as the end of 2020. Unfortunately, individual files are mostly of non-standard format requiring a significant effort in cleaning and standardizing these data sets. The rapid growth of data will create significant challenges in gathering, cleaning, standardizing, managing, and exchanging the data.

Our results indicate that accurate SCT classification can be made using ANN prediction models. Although the major cell subtypes can be determined by a small number of cell surface expression markers in cell sorting studies, these markers are often not captured in SCT data, and often subsets of different cell subtypes express overlapping sets of surface markers. We have shown that supervised machine learning can compensate for both limitations in measurements and biological patterns overlap. In practice, this allows us to skip the cell sorting step and directly analyze mixed PBMC samples.

Machine learning methods involve optimization of performance. Increasing the number and quality of training data sets and generating high-quality test sets is the basic approach. More advanced methods include feature extraction and dimensionality reduction, optimization of model architecture and learning algorithms, exploration of multiple machine learning algorithms, and the use of knowledge-based methods. The availability of large number of standardized SCT data sets has enabled the application of supervised machine learning methods, paving the way for development of new SCT-based blood tests.

CHAPTER 5 STUDY II - INCREMENTAL LEARNING

Systematically incremental learning experiment design and cyclical validation on SCT PBMC classification have been deployed for ANN model training and testing in this study. This work has been organized and published on the 2020 International Conference on Bioinformatics and Biomedicine (BIBM) [65].

5.1 Abstract

In this study, we obtained and standardized 27 SCT data sets, derived from healthy PBMC samples using 10x SCT. We used artificial neural networks (ANN) to assess the ability of ANN to classify main PBMC cell types. Incremental learning by the gradual addition of new data sets to ANN training improved classification. The overall prediction accuracy of the final step of incremental learning reached 93% in 4-class classification.

5.2 Introduction

Supervised learning methods, such as artificial neural networks (ANN), can be used for advanced SCT cell classification with the potential for automation of analysis. Previously we standardized a selection of PBMC data sets and applied artificial neural networks (ANN) to explore its ability to classify main cell types of PBMC. We achieved the accuracy of five-class classification of human peripheral blood mononuclear cells (PBMC) to be approximately 90% [16]. In the current study, we extended the previous model to a full, incremental learning model to classify 5 main cell types of PBMC. Three research questions were pursued in this study:

- Can incremental learning (retrain ANN with newly generated data) improve the accuracy of classification?

- Can this classification system learn by combining data from samples that are subject to very different sample processing methods?

- How stable is ANN model performance as new independent data sets are added?

5.3 Materials and Methods

5.3.1 Study design

We deployed incremental learning (data accumulation methodology [253]) for ANN model training and testing. The design aims to study the data quality effect to single cell classification performance, as simulating the real-life situation – when new diverse SCT data sets are generated from different laboratories/hospitals and added into the previously existing training data set. In each cycle, 2-fold cross validation, external validation with the next upcoming data set, external validation with a qualified test data set (BroadS1 data sets), have been conducted to evaluate the trained ANN model. At the end of this cycle, the next upcoming data set is added into the existing training set and forms a new accumulated training set. In the next cycle, this newly generated accumulated training set is used to train the ANN model, and the same validation steps are repeated as the last cycle. In each cycle, the performance assessment is done with determined metrics, as described in Methodology Chapter, for five cell types of PBMC.

The training data consisted of the 10x Gen data sets [10] and GEO DB data sets [251], derived from multiple independent studies. The training and testing of ANN consisted of several iterated cycles where training was done using continuously increasing independent multi-source data sets. Nine 10x Gen data representing four cell classes (B cells, monocytes, NK cells, and T cells) were used as the initial training data set (the first cycle, Table 9). Thirteen GEO DB data sets were ordered based on study publication date and used in cycles 2, 3, and 4 as shown in (Table 9). Since our training data did not have a dendritic cell set, the ANN predictor was trained as a 4-class classifier. Overall, our study had 25 training-testing steps distributed over five training cycles.

Each training-testing cycle had three parts: internal cross-validation (2-fold), classification of new incoming data sets, and external validation. The classification of new data sets was performed using ANN models trained by all data sets available in the immediate previous cycle. BroadS1 data set was used as a test set for external validation (ICA dataset, singlecell.broadinstitute.org). We consider it as a suitable testing data set since it was checked and annotated by experts. BroadS1 has a class DC with 142 instances of dendritic cells. Because we did not have DC in the training sets, we merged DC from BroadS1 into monocyte test set.

The flow chart describing the design of this study is shown in Figure 36. The loop in the middle of the chart was repeated for each of the 25 steps in our study. The data sets were added to the training set ordered by the date of their addition to the GEO DB.

5.3.2 Data

We collected, cleaned, and converted into standard format 27 SCT data sets of PBMC. These data sets were generated from fresh and frozen blood samples using 10x sequencing technology. Nine datasets were from 10x Gen; 13 datasets from 5 GEO studies (GSE103544, GSE112845, GSE116130, GSE116683, and GSE124731). The BroadS1 dataset from study ID SCP345 was used for the test set. The number of cells used in this study is shown in Table 10. Each individual data set in this study was in the form of sparse matrix, having 30,698 rows representing human genes, and up to 11,954 columns representing single cells. In each matrix the number of columns was identical to the number of cells in each dataset.

Table 9. The training set and testing set in each cycle of ANN incremental learning experimental design. Step 26 is added to indicate future inclusions of new data sets.

Cycle	Step	Action	Training sets	Testing sets	Cell type
Cycle 0	Step 1	Cross validation	10x dataset	10x dataset	
	Step 2	Classification	10x dataset	MC0001	CD14+ Monocytes
	Step 3	Classification	10x dataset	MC0002	CD14+ Monocytes
	Step 4	Classification	10x dataset	BroadS1	
Cycle 1	Step 5	Cross validation	nTRS170915	nTRS170915	
	Step 6	Classification	nTRS170915	nTC0101	CD8+ cells
	Step 7	Classification	nTRS170915	BroadS1	
Cycle 2	Step 8	Cross validation	nTRS180725	nTRS180725	
	Step 9	Classification	nTRS180725	BC0201	CD19+ cells
	Step 10	Classification	nTRS180725	BroadS1	
Cycle 3	Step 11	Cross validation	nTRS181015	nTRS181015	
	Step 12	Classification	nTRS181015	NK0301	NK cells
	Step 13	Classification	nTRS181015	TC0302	CD4+ T cells
	Step 14	Classification	nTRS181015	TC0303	CD8+ T cells
	Step 15	Classification	nTRS181015	TC0304	iNKT (invariant Natural Killer T cells)
	Step 16	Classification	nTRS181015	TC0305	MAIT (Mucosal-associated Invariant T cells)
	Step 17	Classification	nTRS181015	TC0306	Gamma Delta 1 T cells
	Step 18	Classification	nTRS181015	TC0307	Gamma Delta 2 T cells
	Step 19	Classification	nTRS181015	BroadS1	
Cycle 4	Step 20	Cross validation	nTRS190108	nTRS190108	
	Step 21	Classification	nTRS190108	TC0408	CD4+ T cells
	Step 22	Classification	nTRS190108	TC0409	CD4+, CCR5+ CD69- T cells
	Step 23	Classification	nTRS190108	BroadS1	
Cycle 5	Step 24	Cross validation	nTRS190620	nTRS190620	
	Step 25	Classification	nTRS190620	BroadS1	
	Step 26	Classification	nTRS190620	… …	

Table 10. Total number of cells for different cell types and data sources implemented in this study.

Cell type/ Total number of cells	10x Gen	GEO DB	BroadS1	Total
B cells	10,085	1,760	1,751	13,596
Dendritic cells	0	0	142	142
Monocytes	2,612	856	1,668	5,136
NK cells	8,385	309	1,394	10,088
T cells	64,347	8,789	8,344	81,480
Total	**85,429**	**11,714**	**13,299**	**110,442**

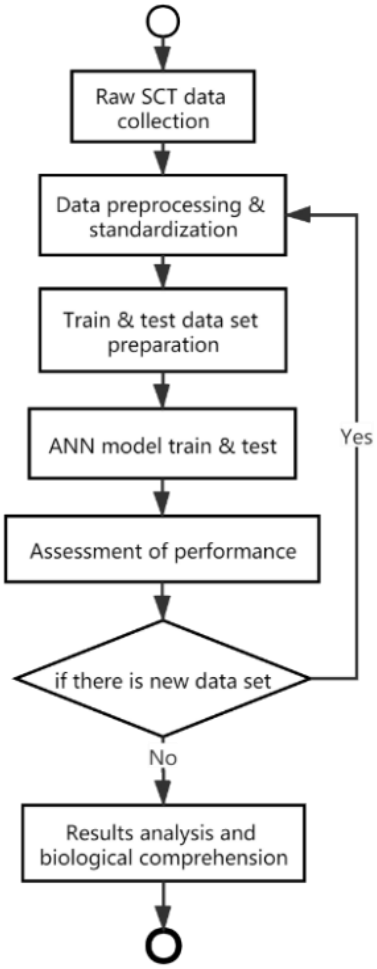


Figure 36. Experimental design with incremental learning for ANN classification of PBMC cell types using SCT data.

5.4 Results

ANN classification of 10x SCT data sets from healthy PBMC samples was done using incremental learning using independent data sets. We analyzed the change of accuracy of incremental learning in each step on specific cell types. Then, we assessed the overall accuracy at the end of each cycle. Finally, we assessed the performance of ANN classifier on specific cell types by considering all performance measures.

5.4.1 Incremental learning

During the incremental learning, the initial ANN was trained by a combined data set composed of nine 10x Gen data sets (B cells, monocytes, NK cells and six T cell data sets). Thirteen SCT data sets of healthy PBMC samples from GEO database were adding for incremental learning in order: M →M→T→B→NK→T→T→T→T→T→T→T→T, where B, M, NK, and T stand for B cells, monocytes, NK cells, and T cells, respectively. The results (Figure 37) show that the initial ANN trained on 10x Gen data could predict NK cells with high accuracy and T cells with low accuracy (50%), while the accuracy of classification of B cells (73%) and monocytes (85%) was intermediate (Step 4, Figure 37). Adding monocytes to the training data increased the accuracy of classification for monocytes while accuracy of classification of other cell types decreased slightly (Step 7, Figure 37). Adding one T cell data set resulted in a notable increase in the accuracy of T cells (from 47% to 92%), while the accuracy of NK cells decreased (from 96% to 80%) (Step 10, Figure 37). Adding one NK data sets to training (Step 19, Figure 37), stabilized prediction accuracies to be close to 90%. Adding multiple T cells stabilized the accuracy of classification of B cells (90%) and monocytes (99%), and T cells (97%), while it did affect the accuracy of classification of NK cells. The final accuracy of NK cells reached 73% (Step 25, Figure 37).

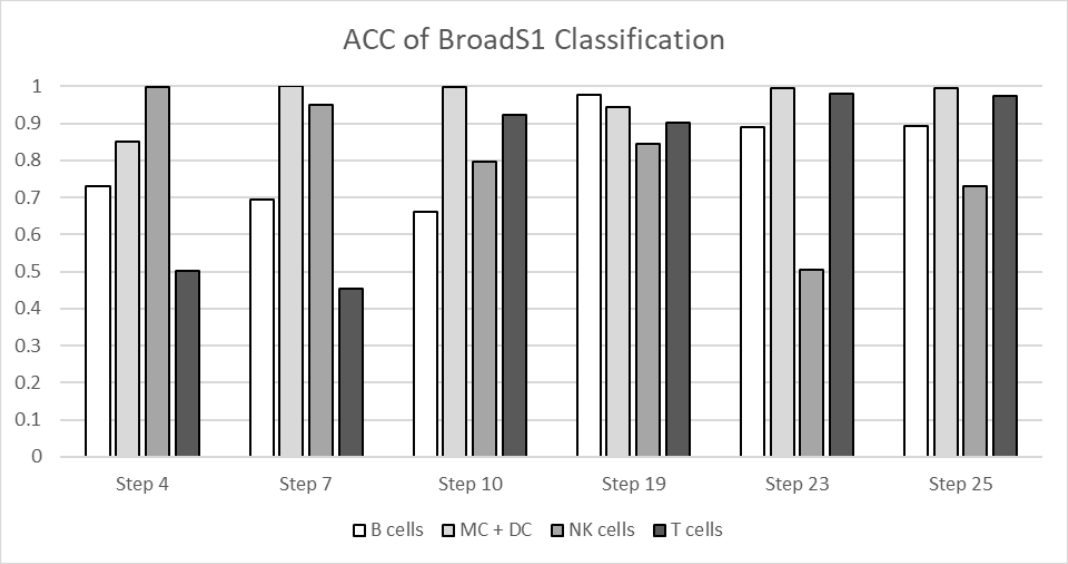


Figure 37. ANN performance on cell type classification of the incremental learning experiment across different cycle steps.

5.4.2 Overall accuracy

The overall average classification accuracy of B cells, MC+DC, NK cells, and T cells showed steady improvement as the training set was increasing (Figure 38). The exception was a slight decline in overall accuracy in step 7. The overall average of all these cell types across all the steps in incremental learning procedure has grown from 0.62 to 0.93, from step 4 to the final step 25.

We used micro-average method to calculate the average value. Micro-average (total true prediction/total number) weighs each sample equally whereas macro method weighs each class equally. In our multi-class classification setup, micro-average is preferable when there is class imbalance (considering DC class and TC class).

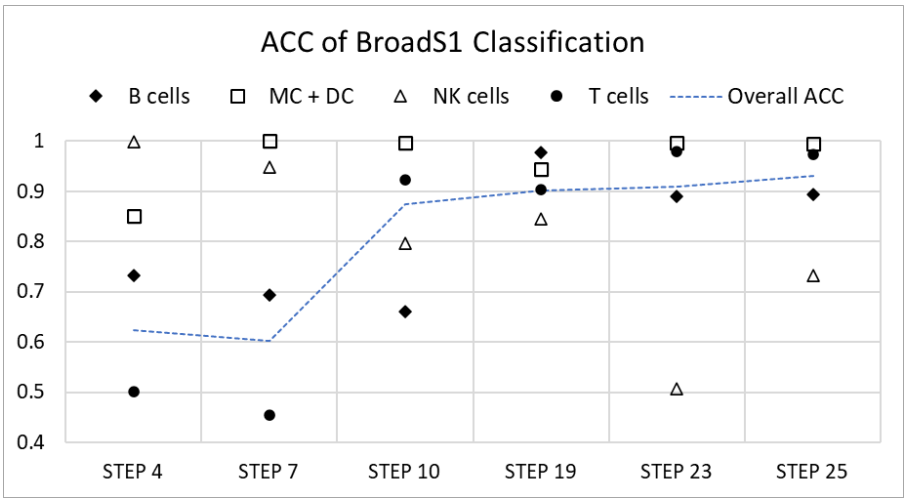


Figure 38. The overall accuracy of the classification of ANNs during incremental learning across different cycles. Data sets were added in order following study publication dates, from earliest to the latest.

The ANN model trained incrementally shows a steady improvement of the overall accuracy. However, we can observe a lack of stability of accurate predictions for specific types of cells. Adding a data set to training can markedly change predictions. For example, extensive changes were seen between steps 10 and 19 (Figure 38). Adding a NK data set to training data increased accuracy of B cell classification from 67% to 97%, and of NK cells classification from 79% to 84%. On the other hand, the accuracy of classification of monocytes declined from 99% to 94% and of T cells from 92% to 90%. Adding multiple sets of T cells may cause changes in the accuracy of NK cell classification (steps 23 and 25, Figure 38).

5.4.3 Sensitivity and specificity analysis

The SE/SP analysis tells us about positive prediction rates and negative prediction rates. The results (Figure 39) show satisfactory predictions for monocytes. Classification of B cells shows high specificity and sensitivity of ~90%. This means if a vast majority of cells predicted as B cells are indeed B cells. On the other hand, 10% of actual B cells will be classified as some other cell type. Another important observation is that we have a notable bilateral misclassification of T cells and NK cells. We propose that this misclassification involves NK-like T cells [254].

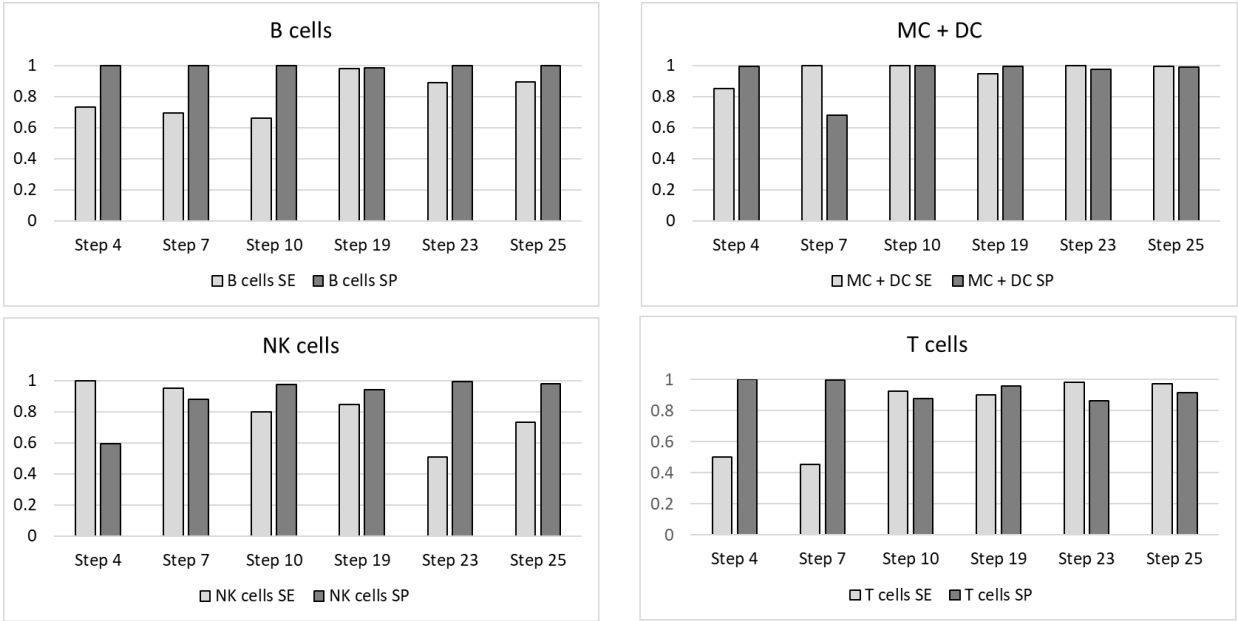


Figure 39. ANN predication performance on each cell type in the incremental learning experiment.

5.4.4 Final step results

The overall accuracy of the final step predictions reached Acc=93.0% (Table 11). In step 25, 10.6% of B cells, 0.54% of monocytes, 26.8% of NK cells, and 2.6% of T cells were misclassified. The highest misclassification rate was for NK cells – 26.5% of experimental NK cells were classified as T cells. The second highest misclassification was for B cells – 6.3% of experimental B cells were classified as monocytes, and 2.9% as T cells. 2.4% of experimental T cells were classified as NK cells.

Table 11. The confusion matrix of final training and testing cycle (step 25).

Experimental \ Predicted	B_cells	Monocytes	NK_cells	T_cells	Sum
B_cells	**1,565**	111	25	50	1,751
Dendritic_cells	0	**142**	0	0	142
Monocytes	0	**1,659**	2	7	1,668
NK_cells	1	3	**1,021**	369	1,394
T_cells	10	10	201	**8,123**	8,344
Sum	1,576	1,925	1,249	8,549	**13,299**

These results were corroborated by the PR/RE and F1 classification performance metrics (Table 12).

Table 12. The assessment metrics of the final training and testing cycle (step 25).

	B_cells	MC+DC	NK_cells	T_cells
Precision	0.993	0.862	0.817	0.950
Recall/Sensitivity	0.894	0.995	0.732	0.974
Specificity	0.999	0.977	0.981	0.914
F1_Score	0.941	0.923	0.773	0.962
Accuracy		0.930		

5.5 Conclusions and discussion

Compared to the previous work [16], we used additional data sets and excluded several data sets that do not represent healthy PBMC. The incremental learning demonstrated the overall accuracy improvement from 89% to 93%. Gradual but steady improvement of the overall accuracy indicates that the overall strategy is successful, and future improvements will be achieved by the addition of new data sets. The addition of new data, however, needs to be done with due care. We observed that new data sets could cause marked shifts of misclassifications from one class of cells to another. We observed the bilateral misclassifications within the B cells-monocytes and NK cell-T cell pairs.

An important observation from our study is that the training data and test data do not represent the same sample processing steps. Our training data involve more processing steps than the test set, since training data involve cell sorting by FACS instrument while the test set was annotated by feature analysis and expert annotation. This indicates that although additional sample processing steps do change gene expression profiles, the fundamental patterns of gene expression remain preserved in the cells, thus enabling accurate classification. For bulk sequencing, FACS sorting has minimal effects on gene expression profiles [241]. However, we found that in SCT gene expression profiles show large differences between gene expression profiles of unsorted cells and profiles of cells sorted by FACS [28]. ANN models showed robustness and the ability to capture key patterns of cell classes irrespective of the sample processing.

There are several limitations of this study that will be addressed in future work. The training data set, although diverse, is limited. We have only two independent data sets of NK cells, two sets of B cells, and three sets of monocytes. Additional data sets are needed to capture the diversity of cell subtypes. We do not have DC in training sets, and these data need to be added. The addition of new data sets must be done with care to prevent large changes in predictions for specific cell types.

CHAPTER 6 STUDY III –INCREMENTAL LEARNING WITH PURIFIED REFERENCE DATA AND FOUR SUPER SETS SWAPPING EXTERNAL VALIDATION

The work of this chapter has been organized and documented into journal paper manuscript.

6.1 Abstract

We used 56 purified reference datasets to train ANN incrementally – over seven cycles of training and testing. The sample processing involved four protocols: separation of PBMC, separation of PBMC + enrichment (by negative selection), separation of PBMC + fluorescence-activated cell sorting (FACS), and separation of PBMC + magnetic-activated cell sorting (MACS). The training data set included between 85 and 110 thousand cells, and the test set had approximately 13 thousand cells. Training and testing were done with various combinations of data sets from four principal data sources. The overall accuracy of classification on independent data sets reached 5-class classification accuracy of 94%. Classification accuracy for B cells, monocytes, and T cells exceeded 95%. Classification accuracy of natural killer (NK) cells was 75% because of the similarity between NK cells and T cell subsets. The accuracy of dendritic cells (DC) was low due to very low numbers of DC in the training sets.

The incremental learning ANN model can accurately classify the main types of PBMC. With the inclusion of more DC and resolving ambiguities between T cell and NK cell gene expression profiles, we will enable high accuracy supervised ML classification of PBMC. We assembled a reference data set for healthy PBMC and demonstrated a proof-of-concept for supervised ANN method in classification of previously unseen SCT data. The classification shows high accuracy, that is consistent across different studies and sample processing methods.

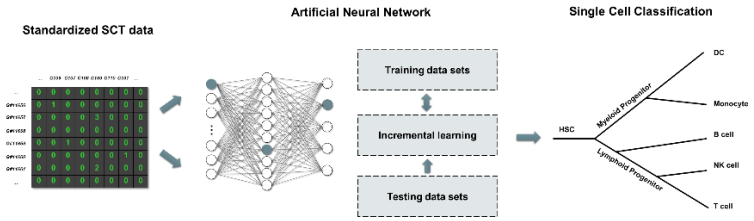


Figure 40. Graphic abstract for Study III. This study is a baseline research to investigate the performance of ANN models with purified reference SCT data.

In this study, we prepared purified SCT datasets to perform incremental learning. Also, the newly collected datasets of BroadS2 were added in the cycles, that brought unseen profiles and training instances for dendritic cell class. In the second part of this study, four data sources swapping external validation experiments has been performed, to investigate the effect of data generating protocols to classification performance.

6.2 Introduction

Our earlier work demonstrated the potential of artificial neural networks (ANN) to classify healthy PBMC cells in blood samples. In the original study, we achieved the accuracy of PBMC classification (BC, DC, MC, NK, and TC) of 89.4% [16]. The follow-up study was performed using an improved and expanded data set to perform incremental learning. Several irrelevant data sets were removed, such as DC from non-blood samples (tonsils and tumor ascites) and T cells fixed in methanol, and several new data sets were added to the training set. The classification accuracy improved to 93.0% [65]. The introduction of assemblies of ANNs with a new voting function further improved the accuracy of classification to 94.7%, but this required a 100-fold increase in computational processing time.

The previous two studies have demonstrated that high accuracy can be achieved in the single cell classification of PBMC cell types. The limitation of these studies is that all testing was performed using a single independent (of the training) test set that was annotated by experts. In this work, we used experimentally labeled datasets to test the trained model. In the current work, we have explored generalization properties of the ANN classification by incremental learning, the effects of data protocols on classification accuracy, and have assessed the current accuracy of PBMC classification by ANN. This study is vital for establishing a baseline for comparing healthy samples with those representing various altered conditions, including gene expression changes in disease.

This study is an extension of our previous studies [16, 65]. The basic ANN classifier is the same as in previous studies. The data sets used for training and testing are different: some of the data sets used in [65] were removed and new data sets were added. Subsequent analysis of data sets used in our previous study indicated that some of the training data represent cells that were processed to the extent that they do not represent healthy PBMC well. The removed data sets include those representing non-malignant cells generated from cancer patients (cutaneous T-cell

lymphoma) pre- and post-therapy (GSM3478792 and GSM3558027) [255], ex vivo activated of T cells (GSM3430548 and GSM3430549) [256], cells that represented mixtures of monocytes and dendritic cells (GSM3258345 and GSM3258347) [257], and cells of mixed populations (selected by designed sorting panel: CD19+ cells (GSM3258348) [257], CD8+ cells (GSM3087628) [49]). One more high-quality test set, BroadS2, was added to our study (GSE132044, [18]).

Compared to former studies, in this study, we added instances representing dendritic cell class into training sets, also brought one more independent data source into the models.

The first part of the study design included incremental learning with larger and more diverse data sets than in our previous studies [16, 65]. The second part of the study involved a comparative validation where all data from one source were used for performance testing while data from other sources were used for training.

Incremental learning is endowed with the ability to continuously process the constantly emerging SCT data, it can retain, integrate, recognize, and extract gene expression pattern of different cell type from accumulated SCT data and newly absorbed data sets.

With multi-source independent data, data accumulation incremental learning can validate the model performance on identifying the effective classification patterns from training knowledge. The accumulation of old knowledge and new knowledge can help the model learn the classification patterns better, and continuously improve the model's ability to make classification judgments. The study has demonstrated the joint training method – traditional data accumulation method for incremental learning. The data accumulation method is to retrain the model on currently all known data. It is generally regarded as - the upper bound of the performance of incremental learning, with the best effect among different learning frameworks. But the disadvantage exists that the training cost is relatively higher.

Cross-validation is added at each training and testing step. The design has discussed how the publication date, batch effect, sampling protocol, and other influencing factors affect the ANN model's ability/behavior to classify the five cell types of PBMC. At the same time, the behavior of ANN classifier on recognizing dendritic cell expression pattern has been discovered.

This study tries to explore four research questions:

• What is the best accuracy of ANN trained using scRNA-seq data to classify five main classes of PBMC?

• How does using data from different studies using different levels of sample processing affect the

accuracy of single cell classification by ANN?

• What is the accuracy of classification when the ANN is trained using samples that have same processing level but are from different studies?

• What are the effects of technical noise on the accuracy of ANN classification?

6.3 Materials and Methods

6.3.1 Study design

In the first part of this study, we deployed an incremental learning process for ANN model training and testing as previously described [65]. Five data sets from BroadS1 study were combined to be used as the test set. The training was performed incrementally – data were added to the training set following the order of time of data sets acquisition. Seven cycles of training were done until all training data sets were used. The overall assessment of classification performance was done after Cycle 7. In the final step, we swapped BroadS1 and BroadS2 data sets and assessed the classification accuracy with BroadS2 dataset as a test set. The incremental learning process is illustrated in Figure 41.

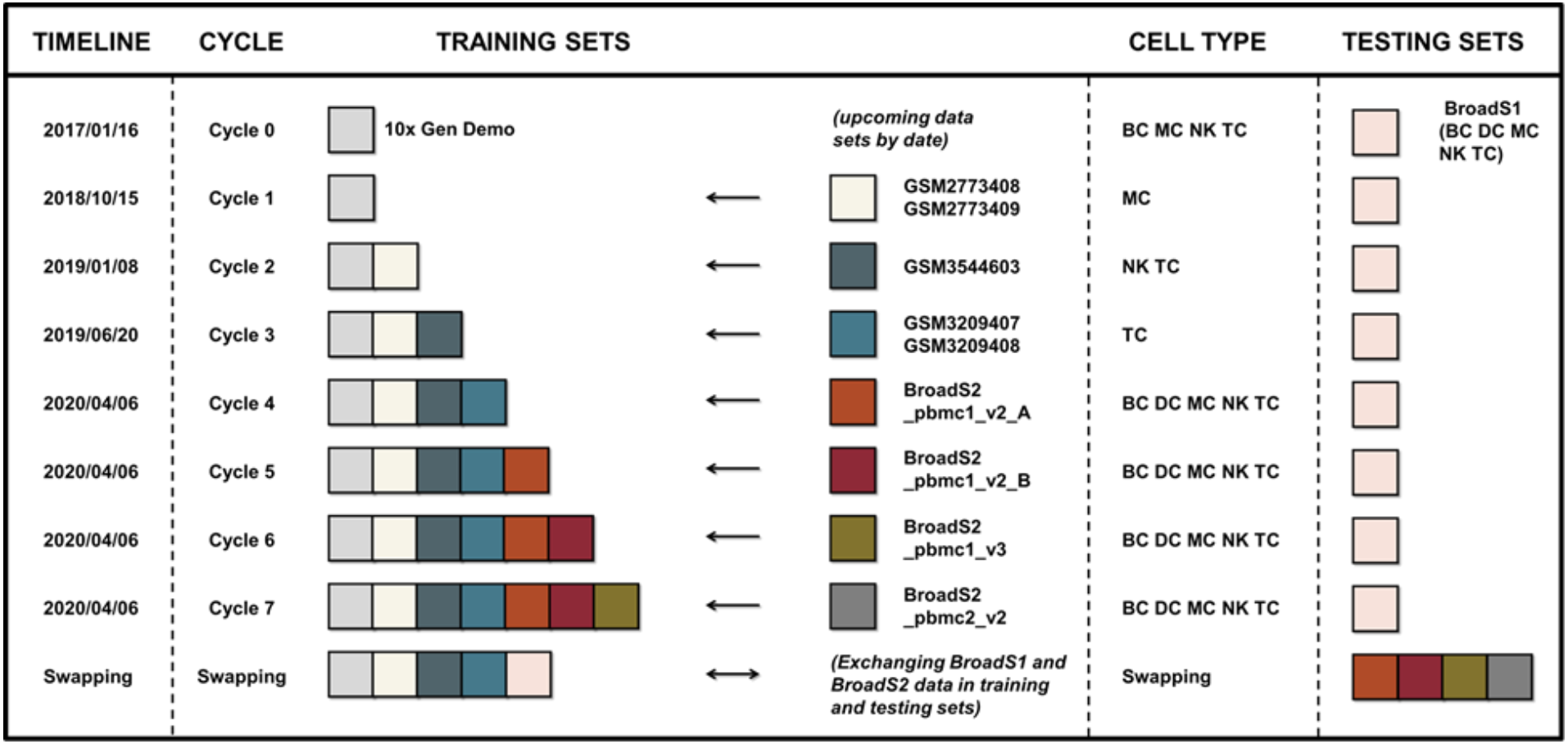


Figure 41. Illustration of the process of incremental learning (training and testing) by adding data sets to the training set and cyclical assessment of classification accuracy. The cycles of learning were ordered by their publication dates to simulate the situation with real-life data accumulation. In the final step of incremental learning, BroadS1 and BroadS2 datasets were swapped to observe the reproducibility of ANN results.

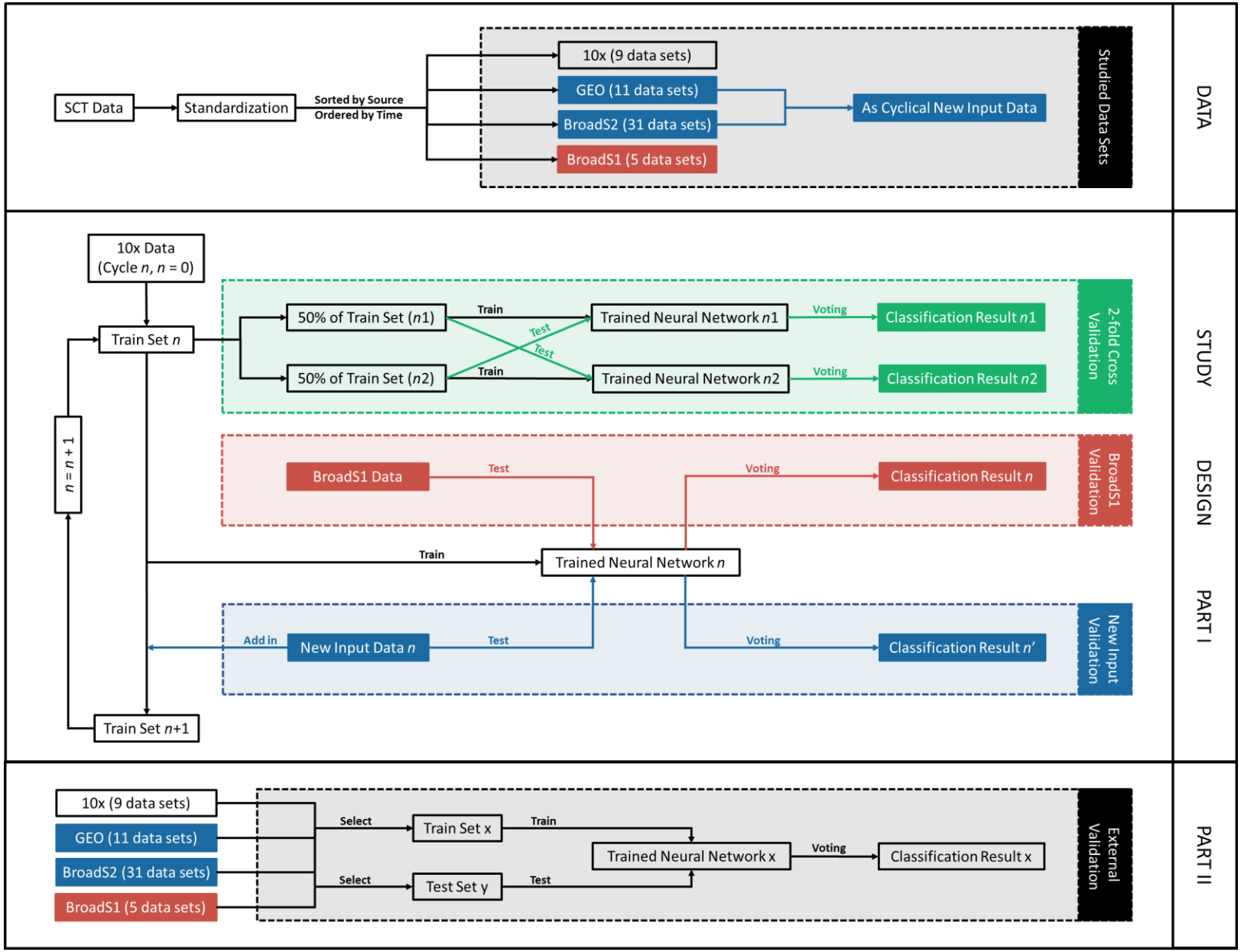


Figure 42. Technical route diagram for the study design in Study III. As illustrated, the study design includes two parts. The detailed is documented as following.

Three types of classification tests were performed in each learning cycle (except Cycle 0 and the Swapping Cycle, that do not have upcoming data sets):

- Internal 2-fold cross validation on the training set to check the internal consistency of the training data,

- Classification accuracy on newly added data sets (upcoming data) before their inclusion in the training set, to check to what extent the gene expression patterns of the added data sets are already represented in the training set,

- Classification accuracy of the training set after inclusion of the added data sets using standard independent test set (BroadS1).

The second part of this study involved a comparative analysis of PBMC classification of different training and testing sets. We performed a comparative analysis of the classification of PBMC using four parallel classification models using data sets from our sources:

- Training set: {10x ∪ GEO ∪ BroadS2}, testing set: {BroadS1}

- Training set: {10x ∪ GEO ∪ BroadS1}, testing set: {BroadS2}

- Training set: {GEO ∪ BroadS1 ∪ BroadS2}, testing set: {10x}

- Training set: {10x ∪ BroadS1 ∪ BroadS2}, testing set: {GEO}

The comparative analysis involved the assessment of classification accuracy and the interpretation of results using the statistical properties of the data sets. A schematic diagram of the detailed overall experimental design is shown as Figure 42.

The model training and testing steps were performed as illustrated. The voting results of the trained neural networks were collected and analyzed of each step, in study part I and part II (in Figure 42).

6.3.2 Data

We selected 56 purified reference datasets that represent PBMC from healthy blood samples. These data sets were collected from the NCBI GEO database (www.ncbi.nlm.nih.gov/geo), 10x Genomics demonstration data [10], and Broad Institute Single Cell Portal (singlecell.broadinstitute.org/single_cell). All data sets were processed into our standardized format that has 30,698 features (genes). Most analyses were done using raw data values of standardized features, as provided by the source. Additional validation step was performed with cells from BroadS2 that were subject to quality control: cells that had less than 300 positive features, or less than 670 total counts were excluded, and the results were compared to the results obtained from predictions that used raw data only. 10x demonstration data were generated using standardized 10x scRNA-seq experimental protocol, including validated upstream data analysis [10]. We consider these data sets as reference for PBMC cells processed by PBMC isolation, enrichment (purification), freezing, thawing, and 10x processing.

Eleven data sets were extracted from the GEO database including GSM2773408, GSM2773409, GSM3544603 (seven datasets in this GSM), GSM3209407, and GSM3209408 [209, 258, 259]. These data sets were generated from PBMC samples extracted from fresh whole blood of healthy donors. These 11 data sets were produced using 10x protocol after PBMC isolation followed by cell sorting by FACS (fluorescence-activated cell sorting) or MACS (magnetic-activated cell sorting). We obtained two PBMC data sets from Broad Institute Single Cell Portal. The first data set is from the study SCP345, and the second data set is from the study SCP424 (also published in GEO GSE132044 [18]). We named these two data sets BroadS1 and BroadS2, respectively. These data sets were produced using 10x protocol after PBMC isolation followed by annotation of cell types by cell labeling algorithms, and manual labeling correction by experts. These data sets represent a large variety of sample processing, experimental conditions, data analysis approaches, and study purposes. The original test sets (BroadS1) and the newly added set (BroadS2) have multiple repeated SCT measurements of samples from the same individuals at different times, locations, or different chemistry. The same samples processed under the same conditions show high reproducibility. When different chemistry (v2 vs. v3 with BroadS2) was used in the 10x protocol, a modest but notable shift in gene expression reproducibility was observed [28]. The summary information on the distribution of cell types across our data sources and their numbers is shown in Table 13. The detailed description of data sets with associated metadata can be found in Supplemental Table 1 (in Appendix 7 Supplemental Materials in Study III, same as the followings).

The number of data sources in our study is four, and the number of data sets is 56. PBMC comprises five main cell types: B cells (BC), dendritic cells (DC), monocytes (MC), natural killer (NK) cells, and T cells (TC). Cell types in our data set have multiple subtypes: NK cells have one subtype; each of BC, DC, and MC has two cell subtypes; TC type has three cell subtypes (Figure 43). TC subtypes are further divided into three sub-subtypes, each for CD8+ T cells and innate-like T cells, and four sub-subtypes of CD4+ T cells. The actual number of PBMC subtypes at multiple levels of ontology is likely to be in hundreds [163]. The total number of cells in our study is 115,190. The test sets have 13,183 cells (BroadS1) or 12,292 cells (BroadS2). The distributions of gene expression values across cells in each data set were visualized. Plotting module pl.violin from SCANPY [124] was used for drawing violin plots.

Table 13. Summary description of 56 SCT data sets involved in this study. Cell numbers and the number of data sets (values within brackets) are shown per cell type. The data sources are described in the main text. BC – B cells, DC – dendritic cells, MC – monocytes, NK – natural killer cells, TC – T cells. BroadS1 is the original test set.

	CELL TYPES - CLASSES					
SOURCES	**BC**	**DC**	**MC**	**NK**	**TC**	**TOTAL CELLS**
10x Demo	10,085 (1)	0	2,612 (1)	8,385 (1)	64,341 (6)	85,423 (9)
GEO	0	0	856 (2)	309 (1)	3,127 (8)	4,292 (11)
BroadS1	1,660 (1)	142 (1)	1,661 (1)	1,394 (1)	8,326 (1)	13,183 (5)
BroadS2	1,884 (4)	270 (7)	2,132 (8)	842 (4)	7,164 (8)	12,292 (31)
TOTAL	13,629 (6)	412 (8)	7,261 (12)	10,930 (7)	82,958 (23)	115,190 (56)

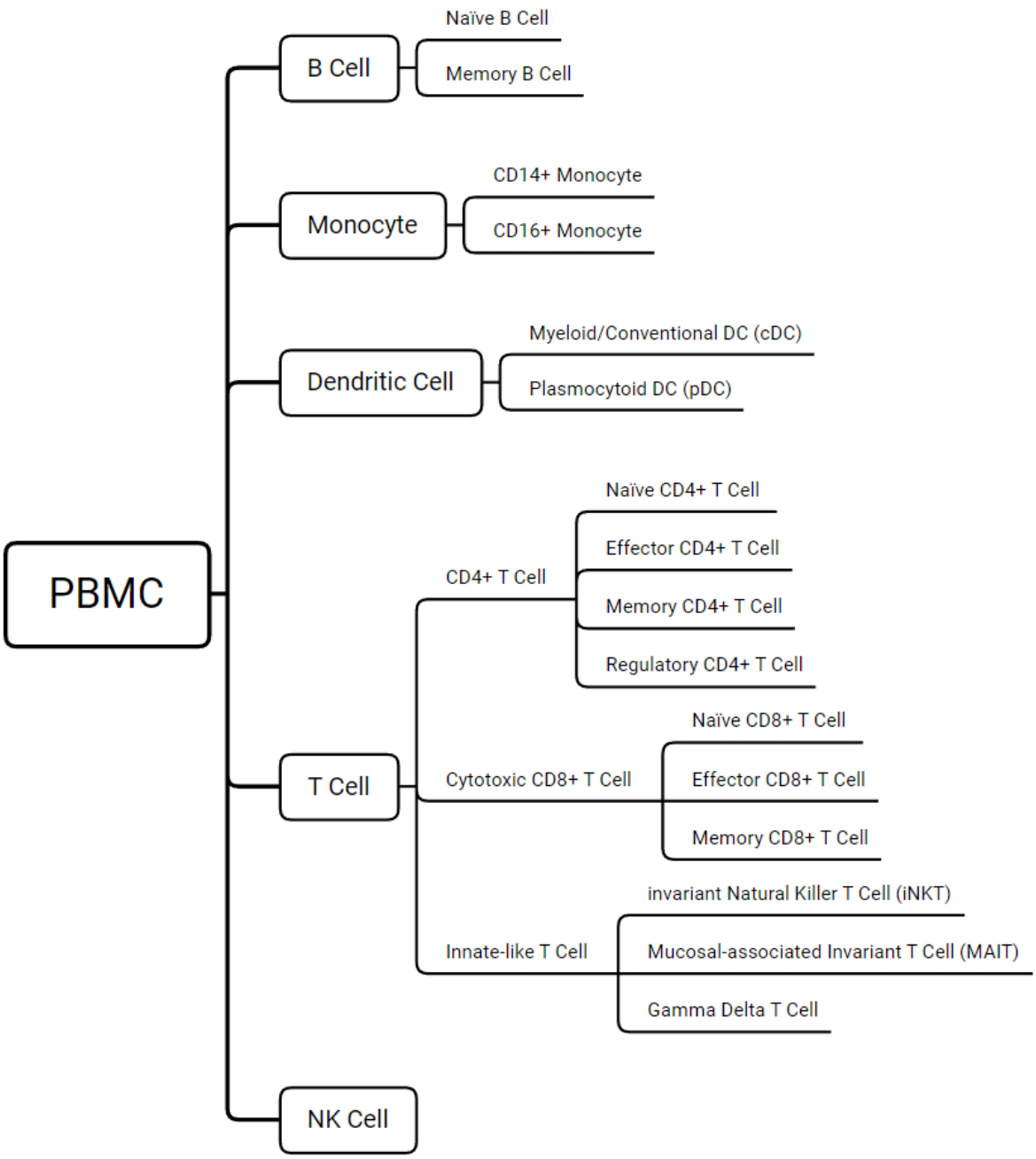


Figure 43. The ontology of cell types and subtypes in our study. The designation of cell subtypes and sub-sub types is provided to show the diversity of cell subtypes used in this study. Because the classification task in this work focuses on the classification of five main types, the descriptions of cell subtypes and sub-sub types have been omitted.

The data are represented as sparse matrices, where the list of cell identifiers (cell ID) occupies the top row (starts from column 2), and the list of gene names (features) occupies the first column (starts from row 2). The first matrix position (1,1) is blank, while other matrix values represent gene expression counts of a given gene in the given cell determined by the matrix position (gene name, cell ID). Our standardized gene list contains 30,698 genes that are arranged in the same order. Most of the values in an expression matrix are zero.

6.4 Results

6.4.1 Density distribution

Density distributions of gene expression within data sets showed a great variety (Figure 44). Data sets from GEO (cells sorted by FACS) show a high median gene expression value (between 2700 and 3300 counts). GEO data sets MC02 and MC03 showed broad quartile ranges and unimodal density distributions. GEO data sets TC13 and TC14 showed intermediate quartile ranges with bimodal distributions. On the other hand, GEO data sets NK02, and TC07-TC12 showed high median gene expression values (around 3000 counts) and narrow quartile ranges, most with bimodal density distributions. NK02 and TC07 data sets showed unimodal distributions and narrow quartile ranges. Bimodal distributions indicate the presence of more than one cell subpopulation.

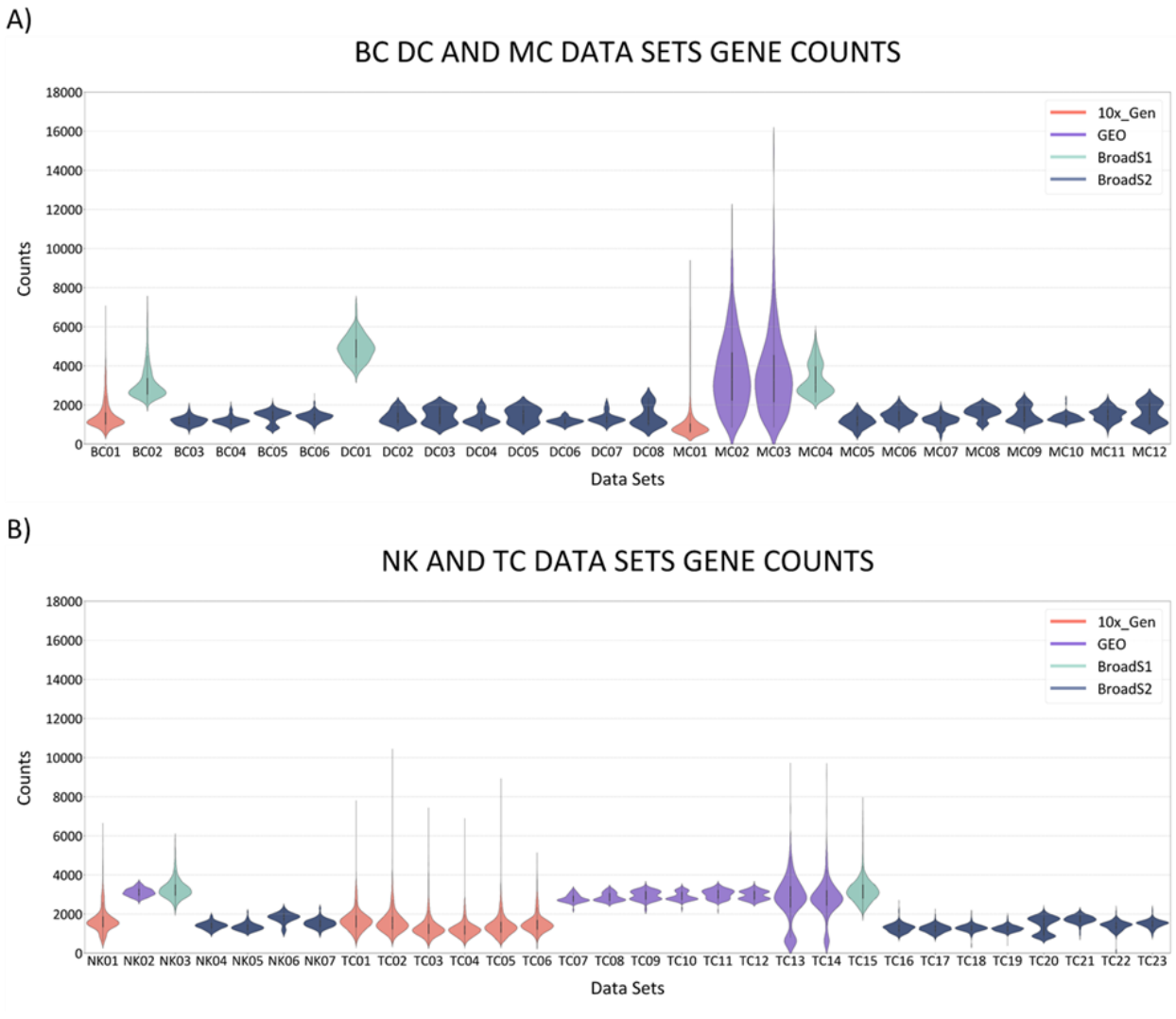


Figure 44. Density distributions of gene expression across 56 data sets used in the current study. A) violin plots of B cells, monocytes, and dendritic cells. B) violin plots of NK cells and T cells. BC01, MC01, NK01, and TC01-TC06 are from 10x demonstration data; MC02, MC03, NK02, and TC07-TC14 are from GEO data set; BC02, DC01, MC04, NK03, and TC15 are from BroadS1; the remaining data sets BC03-BC06, DC02-DC08, MC05-MC12, NK04-NK07, and TC16-TC23 are from BroadS2. The maximal width of each of the violin plots was set to one ("1").

Data sets from BroadS1, BC02, DC01, MC04, NK03, and TC15 showed a high median value of gene expression and intermediate breadth of quartile ranges. The majority of BroadS1 cell type data sets showed unimodal distribution, while MC04 showed a bimodal distribution, most likely representing CD14+ and CD16+ monocyte subtypes. We noted that all BroadS1 data have high gene expression counts (4880 ≥ median counts ≥ 2815, across BroadS1 data sets), and high number of positive features (1890 ≥ median features ≥ 790) than BroadS2 where expression counts (1843 ≥ median counts ≥1163, across BroadS2 data sets) and positive features (1508 ≥ median features ≥ 611) (Supplemental Table 2). BroadS2 data sets showed wider interquartile ranges than BroadS1 data sets. A large proportion of BroadS2 data sets had shown distinct bimodal distributions (Figure 44 B), indicating that this data may contain distinct subtypes within the indicated cell type. Bimodal counts of gene expression were also observed in T cell data sets from BroadS2 data set and in monocytes from BroadS1.

6.4.2 Incremental learning

The average composition of the training sets and the compositions of test sets are shown in Table 14. The composition of the training sets is stable across cycles (Figure 45). Test sets match well the healthy ranges [260, 261] while DC was severely underrepresented in the training sets, monocytes were underrepresented, and T cells were overrepresented (Table 14). The DC were included in the training set only in Cycles 4-7 and their representation in the training set remained low, approximately 10- to 20-fold lower than their representation in test sets. The training set in Cycle 0 included only samples that were from 10x demonstration data – processing of these cells included PBMC extraction, purification by bead-enrichment, freezing, thawing, and 10x processing. Cycles 1-3 included the addition of cells sorted by FACS or MACS to the 10x data set. Testing in all cycles was performed using minimally processed (PBMC extraction and freezing) data set BroadS1. The final round, swapping, involved two steps: a) training data set included 10x, GEO, and BroadS2 data, and testing was done using BroadS1 and b) training set included 10x, GEO, and BroadS1 data, and the entire BroadS2 data set was used for testing.

The results of ANN classification are shown in Figure 46. The internal cross-validation showed reproducibly high accuracy ranging from 99.9% to 99.3%. The accuracy of classification of new independent data sets was initially low (82.0% in Cycle 0 and 24.3% in Cycle 1, then it rapidly increased and stabilized between 92% and 99% from Cycle 2. The external validation with BroadS1 data set showed low accuracy of classification in Cycles 0 and 1, followed by a rapid increase to 92.2%, followed by a gradual improvement in accuracy that reached 94.6% in Cycle 7.

The swapping step, where BroadS2 was used as a test set showed the accuracy of internal cross-validation of 99.2% and external validation accuracy of 91.7%. Taken together, the results indicate that the overall accuracy of 5-class classification is between 92 and 94%.

Table 14. The cell type compositions of training and testing sets. The proportions of the main PBMC cell types are shown for the healthy range [260, 261], training sets, and test sets (BroadS1 and BroadS2).

CELL TYPE	Healthy Range	Average Training Sets	Test Set BroadS1	Test Set BroadS2
B Cells	5-15%	11.44±0.36%	12.59%	15.33%
Dendritic Cells	1-2%	0.09±0.18%	1.08%	2.20%
Monocytes	10-30%	4.44±1.05%	12.60%	17.34%
NK Cells	5-10%	9.64±0.21%	10.57%	6.85%
T Cells	40-70%	74.39±0.93%	63.16%	58.28%

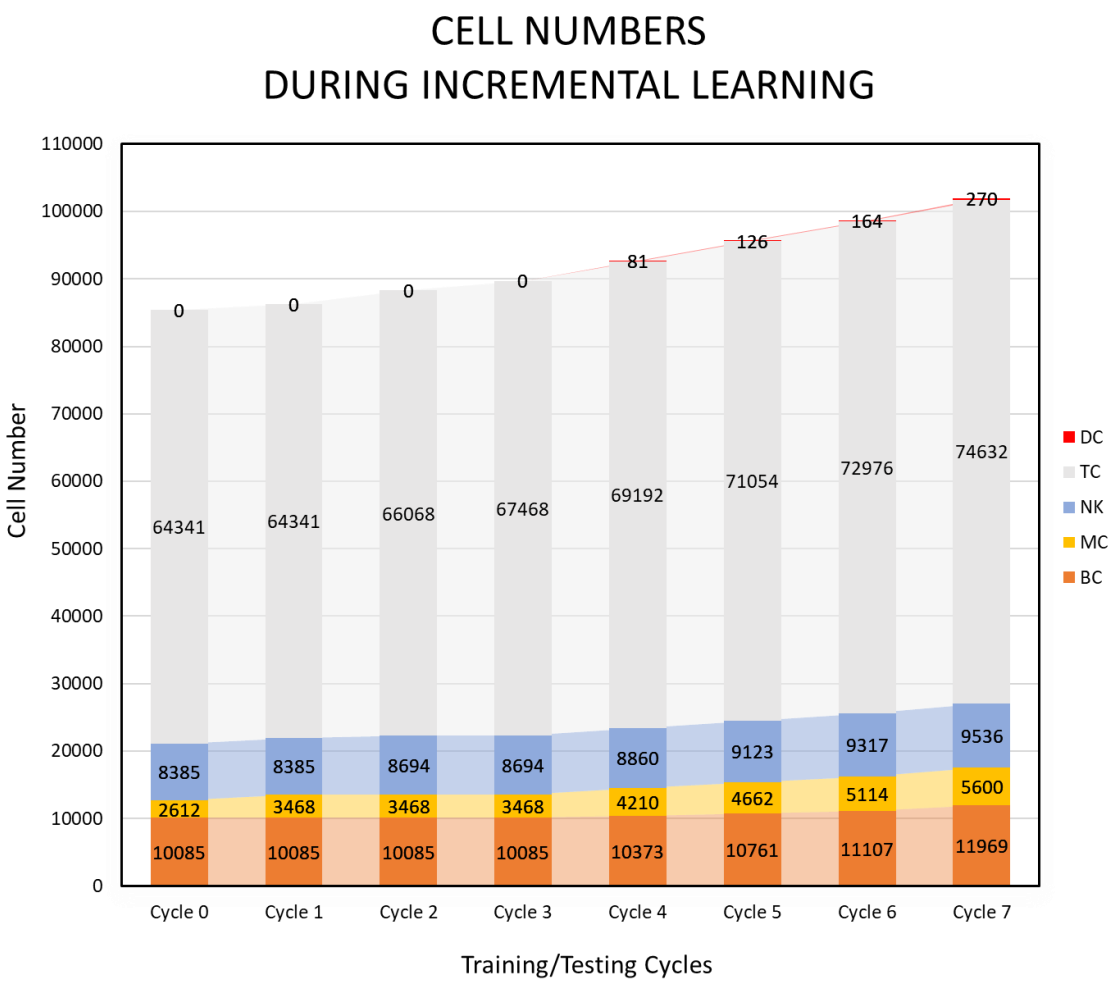


Figure 45. Data sets used in training cycles appear in the time sequence as we acquired them. The increase in the number of cells in training sets was gradual and the proportions of cell types were stable. The new sets of cells tested in the current cycle were appended to the subsequent training set. For example, monocytes from GSM2773408 (425 cells) and GSM2773409 (431 cells) were classified using the training set from 10x dataset (Cycle 0), then were included in the training set for Cycle 1.

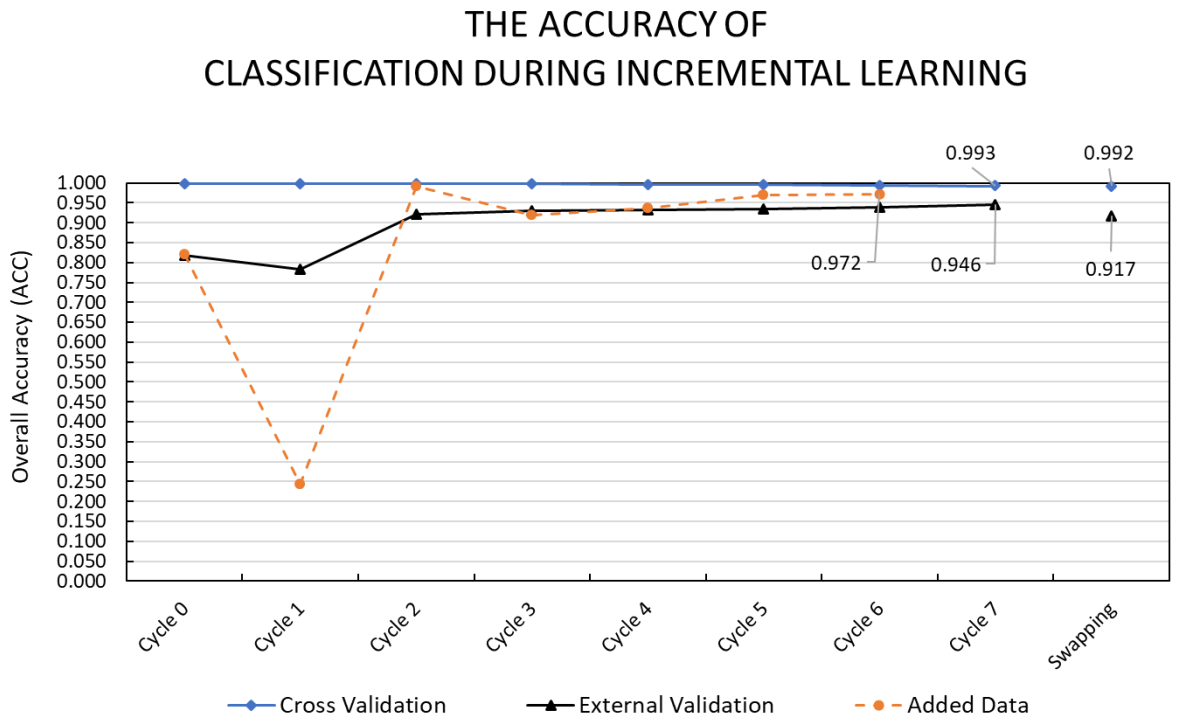


Figure 46. The internal cross-validation showed extremely high accuracy (≥99.2% in all cycles). After early instability (Cycle 1) the classifier starts converging towards the internal cross-validation line. With the increase of the number of data sets added to the training set, new data files are predicted with steadily increasing accuracy (added data line). The swapping step showed that the overall accuracy of the current system is within the range of 92-94%.

6.4.3 External validation

The Cycle 7 and the swapping produced results for EXP 1 and EXP 2 (Figure 46 and Table 15). The remaining part of our study involved training of the ANN classifier by GEO+BroadS1+BroadS2 and testing with 10x data (EXP 3, Table 15), and training of the ANN classifier by 10x+BroadS1+BroadS2 and testing with GEO data (EXP 4, Table 15). Sample processing alone has a profound effect on gene expression pattern recognition (Table 15). The prediction model trained on a combination of samples processed by enrichment or FACS/MACS cell sorting, can be used for high accuracy prediction of minimally processed samples (94.6% and 91.7%, in EXP 1 and 2, Table 15). The model trained with a combination of minimally processed samples can reach higher accuracy, when testing with samples processed by enrichment (98.3%,

EXP3, Table 15) or cell sorting (93.5%, EXP 4, Table 15).

Table 15. Classification accuracy for modeling experiments where the testing set derived entirely from one source, while training sets were combined from other sources. The results also show the F1 measure for individual cell types. Further details are available in Supplemental Tables 3, 4, and 5.

EXP	TEST SET*	SAMPLE PROCESSING	CLASSIFICATION ACCURACY	F1 VALUES
1	BroadS1	Separation	94.6%	BC – 0.963, DC – 0.880, MC – 0.983, NK – 0.781, TC – 0.964
2	BroadS2	Separation	91.7%	BC – 0.962, DC – 0.000, MC – 0.958, NK – 0.695, TC – 0.946
3	10x Demo	Separation, Enrichment	98.3%	BC – 0.969, DC – NA, MC – 0.873, NK – 0.954, TC – 0.995
4	GEO	Separation, FACS or MACS	93.5%	BC – NA, DC – NA, MC – 0.989, NK – 0.700, TC – 0.955
5	BroadS1	Separation	94.5%	BC – 0.953, DC – 0.887, MC – 0.983, NK – 0.792, TC – 0.961
6	BroadS2	Separation	88.1%	BC – 0.876, DC – 0.000, MC – 0.971, NK – 0.592, TC – 0.927

*EXP 1-4 involve three training sets and one testing set. EXP 5 and 6 involve only BroadS1 and BroadS2 data sets.

The overall performance of classification differs between individual cell types (Table 15, EXP 1 and 2): B cells, monocytes, and T cells show high accuracy with F1 values exceeding 0.95. Classification performance of NK cells shows lower accuracy with F1 value in the vicinity of 0.75. Quality control of BroadS2 data set (removal of cells that have total counts lower than 670 or number of positive features lower than 300) did not affect classification performance (EXP 2a and EXP 2b, Supplemental Table 5). Classification of dendritic cells was unstable, F1=0.88 in EXP 1 and 0.00 in EXP 2 (Table 15). When two-fold external validation with BroadS1 and BroadS2 data sets were performed (EXP 5 and 6, respectively), the overall accuracy in EXP 5 was 94.5%, and in EXP 6 was 88.1%. The inclusion of data sets with high median gene expression (GEO, 2700-3300 and BroadS1, 2800-4900, Supplemental Table 2) in the training data set results in lower cell classification accuracy (EXP 2 as compared to EXP 1, and EXP 6 as compared to EXP 5, Table

15). Consistently, adding BroadS1 to the training set in the swapping step, as compared to Cycle 3, results in lower classification accuracy tested on BroadS2 (92.3-91.7%, EXP 7 and EXP 2, Supplemental Tables 3, 4, and 5). ANN model has demonstrated well generalization ability when performing four supersets swapping, it has achieved an average accuracy of 94.5%. Differences in gene expression brought about by various generation protocols have led to differences in predictions for individual cell types, such as the prediction of monocytes was 87.3% in EXP 3 (when trained on a combination of minimally processed samples and samples sorted by FACS/MACS), while in EXP 1, 2, and 3, the monocytes classification accuracy was 98.3%, 95.8%, and 98.9%, respectively (Table 15).

6.5 Conclusions

Overall, our results demonstrate that supervised ML is a viable option for classifying cell types from single cell expression data. Patterns that are characteristic of cell types are preserved in single cell gene expression data even when the single cell samples are processed using different processing steps. Data sets derived from minimally processed samples (PBMC separation only) alone can be used to predict cell type from samples that are additionally processed (we achieved a prediction accuracy of 98.3% for enriched and 93.5% for sorted cells, Table 15). Gene expression pattern characteristics of a given cell type are preserved in samples that have additional processing steps and these sets can be used for accurate predictions of minimally processed samples (93% accuracy on BroadS1 data set was achieved by training set consisting of 10x and GEO data, Figure 46 and Supplemental Tables 3, 4, and 5). That is suitable for broad application. The training data set – the reference set – is composed of multiple data sets that represent various sample processing conditions and contain sufficient biological variability. The ANN classifier is robust – the system can tolerate a proportion of cells that have gene expression lower than quality control thresholds (in our studies it is 670 for gene expression counts and 300 for positive features).

Two-fold internal cross-validation has shown that once a data set is added to the training set, the patterns contained in that set will be remembered by the classifier. The classifier generalizes well, and generalization properties improve with the addition of new data. Once a data set representing a particular cell type and sample processing protocol is added to the training data set, the ANN will learn this data type. When data sets where a particular cell type, biological condition, and experimental processing protocol is well, that is very high (≥99.2%, Figure 46).

The overall classification performances in EXP 1 and 2 (Table 15) are satisfactory (94.6% and

91.7%), also in EXP 3 and 4 (98.3% and 93.5%). Training data used in EXP 1 and EXP 2 are representative of all three sample processing protocols: i. separation (of PBMC), ii. separation + enrichment, and iii. separation + cell sorting. Training data used in EXP 3 did exclude separation + enrichment protocol data that was used for generating test data in the same experiment. Similarly, test data in EXP 4 were generated using separation + cell sorting protocol, while the corresponding training data represented samples produced by other processing protocols. A well-established classification theory concept in ML is that the training set must be representative of the variability that is present in real cases. Our results clearly show the effects of the training sets that are not fully representative. Even the average prediction accuracy of four supersets swapping reaches 94.5%, the effect of enrichment or cell sorting in changing gene expression pattern still appears in the results - when the training set includes data sets of samples by enrichment or cell sorting (EXP 1 and 2), the prediction performance is decreased, compared to when training set includes minimally processed samples (EXP 3 and 4). The data sets of minimally processed samples are found with better representative properties. A problem for SCT is that processing steps such as enrichment or cell sorting are part of the experimental validation of results that are missing in minimally processed samples. Our results of EXP 1 and 2 show high accuracy of classification but cannot be validated directly by experiments. On the other hand, the cell type in EXP 3 and 4 is known, and the classification accuracy are 98.3% and 93.5% when similar data sets are present in the training set.

EXP 1-8 (Supplemental Table 5) results indicate that the average gene expression level of data sets used in training has an influence on classification accuracy, particularly in situations where the training set is limited. The results indicate that the classification of cell types is better in data sets that have moderate gene expression levels, with gene counts between 1000 and 2000 per cell. This observation needs further study to confirm the actual influence of gene counts on classification accuracy. The analysis of factors that possibly influence prediction accuracy in this study is presented in the Discussion section.

In summary, we have demonstrated that ANN, a supervised ML method, is capable of high accuracy classification of five main cell types of healthy PBMC. The accuracy is very high for B cells, monocytes, and T cells. The classification accuracy of NK cells is lower, because of their similarity with subsets of T cells (such as NK-like T cells, subsets of CD8+ T cells, and subsets of innate T cells). This problem was noted in [10], where the authors reported that it was challenging to separate cytotoxic T cells and NK cells since they have overlapping feature spaces. The accuracy of the classification of DC is low because of the underrepresentation of DC in the training sets, and this problem should be overcome by adding additional DC samples.

6.6 Discussion

This work demonstrates the potential of supervised ML methods to classify single cells from their gene expression counts. We achieved the 5-class classification accuracy of 94% using 56 data sets derived from healthy PBMC that were processed by different experimental procedures applied to PBMC samples. An important finding is that once a dataset representative of a cell type, condition (in this case healthy PBMC), and a specific sample processing protocol is added to the training set, similar data sets will be classified with very high accuracy (>99%).

Several factors limit the accuracy of our 5-class classification of PBMC. They include lack of training data (for DC) and similarity of cell subtypes with cells from other classes (NK cells), and training data with high median gene counts. Additional factors include undefined classes or subclasses of cells that are normally found in peripheral blood but are not included in current training set. Such cells, for example, include CD34+ cells (circulating hematopoietic cells that may represent between 0.1 and 0.3% of PBMC [261]. Natural killer T (NKT) cells have markers of both T cells (CD3+) and NK cells (CD56+) and are present in circulating PBMC [262] and can easily be confused with NK cells. On the other hand, CD8+ NK cells [263] share properties with cytotoxic T cells. Given the similarity of gene expression profiles, is not surprising that in our study, 2.6% (218) of T cells from BroadS1 and 8.7% (624) of T cells from BroadS2 were classified as NK cells. Conversely, 22.9% (319) of NK cells from BroadS1 and 6.7% (56) of NK cells from BroadS2 were classified as T cells. FACS sorting showed that NK cells from 10x data were 92% pure, while CD8+ cytotoxic cells were 98% pure. Further investigation, including advanced clustering methods (such as [264, 265]) and the analysis of misclassifications, will be pursued to improve PBMC classification.

One challenge for the classification of cells from SCT data arises from the need for experimental validation of cell types as opposed to expert annotation in minimally processed samples. Experimental sample processing steps such as bead enrichment (negative selection) produce homogeneous samples (one cell type or subtype) whose purity can be verified by cell sorting. Alternatively, cells can be sorted by FACS or MACS procedures that help sort cells, and provide a measurement of purity, percentage of contaminating cells, and cell properties (*e.g.* [258]). Depending on the purpose of single cell study, various sample processing workflows may be deployed (Figure 47). The difficulty with processed samples is that each processing step induces changes in gene expression profiles. These profile changes are significant, and they prevent direct comparison of cells from studies that follow different protocols. Minimally processed samples have similar gene expression to the native blood cells. The annotation of single cells in this case, is done by various tools that utilize gene expression markers and are normally inspected and

corrected manually, introducing annotation bias. Protein markers and gene expression markers do not match perfectly, the expression of proteins and corresponding mRNA significantly correlate only in about one-third of targets [266, 267]. Since SCT data sets are sparse and a large proportion of expressed genes are missing, simple marker-based assignments are insufficient. A selection of *in silico* methods is needed in combination with experimental validation for conclusive assignment of cell types and subtypes.

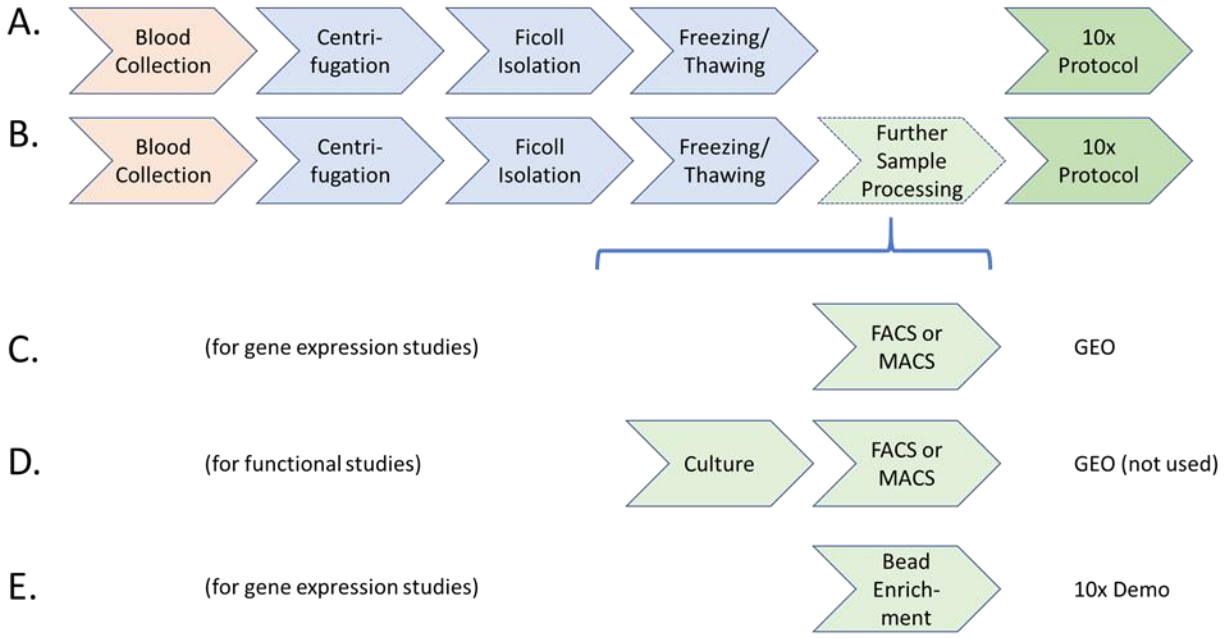


Figure 47. Sample workflows relevant for our study: A. Workflow involving minimally processed samples (BroadS1, BroadS2), B. Generic flow for 10x studies, C. Workflow of samples processed by FACS or MACS, may include multi-step processing (GEO data sets in our study), D. Workflow for functional studies, PBMC samples are often cultured overnight along with bioactive agents, followed by FACS/MACS, E. Workflow using negative selection by bead enrichment (used in 10x demonstration study). Workflow D was not used in this study because culturing with bioactive agents generates cellular responses not relevant for profiling of PBMC from healthy blood.

Supervised ML has distinct advantages in comparison with unsupervised clustering when used for classification tasks. The main advantage is that once reference sets are available, standardized analysis can be performed across samples that represent various biological conditions. Single cell technologies applied to PBMC require the ability to analyze minimally processed samples directly and accurately and reproducibly determine cell types, subtypes, and their status from single cell expression profiles. To achieve this goal, we need standardized sample processing workflows and SOP of upstream single cell analysis and supervised ML methods for downstream analysis. Several sample processing protocols were demonstrated as reproducible and are available (see support.10xgenomics.com/single-cell-gene-expression/sample-prep). SCT samples can be analyzed using existing SOPs and they yield highly reproducible results (as demonstrated, for example, in [18, 28]).

Given that the SCT part is stable, supervised ML requires that training data are representative of all major sample processing protocols. Supervised ML analysis can classify any future sample collected, prepared, and analyzed using one of the validated protocols with the expected accuracy. Our results indicate that the accuracy of classification from validated protocols should be above 98%, which matches cell purity from standard cell sorting methods. New sample processing protocols can be validated by splitting minimally processed samples, perform supervised method (such as ANN) classification on one partition of the sample, and performing additional processing steps to confirm the numbers or proportions of cell types in the second partition. In this study we have defined a reference data set for 10x PBMC 5-class classification that provides 94% accurate classification. Our future goal is to refine classification of DC, by increasing the number of DC data in the training set and resolve ambiguities between NK cells and subsets of T cells (non-classical T cells and CD8+ T cells) that are misclassified due to their gene expression similarity.

CHAPTER 7 STUDY IV - VULNERABILITY OF ANN-SCT-PBMC CLASSIFIERS

In this chapter, we studied the vulnerability of ANN-SCT-PBMC models, using five groups of non-representative datasets and seventeen rounds of 4-supersets-swapping external validation.

7.1 Abstract

The vulnerability and robustness of the ANN-SCT-PBMC model can be affected by SCT data representativeness. This study aims to verify the vulnerability and robustness of the ANN-SCT-PBMC model under the cumulative impact of five confounding factors: 'empty cells', 'other tissue', 'dead cells', 'activated cells', and 'mixed population'. We used 56 reference datasets and 17 non-representative datasets from four independent data sources for deploying 17 rounds of four parallel external cross-validation experiments, to study the classification performance of the model.

The overall average accuracy of four parallel external validation (among 10x, GEO, BroadS1, and BroadS2) increased from 0.660 to 0.945 in 17 train-test rounds when cumulatively eliminating non-representative datasets. The prediction on BroadS1 and BroadS2 testing sets showed high accuracy (averagely 0.937 and 0.914 in 17 rounds). The GEO testing set showed an overall upward trend, it increased with 24.41% of accuracy. The accuracy of the 10x testing set had significant improvement, from 0.059 in Round 1 to 0.983 in Round 17. The performance for four testing sets all converged to above 0.917 at the last swapping round. From the F1-score of each class, BC, MC, and TC prediction was robust, the prediction of NK had lower performance, while the prediction of rare class DC was unstable and affected largely. From the error rate of each cell subtype, misclassification mainly occurred in 'NK', 'nonT', 'DC', 'pDC', four innate-like T cell subtypes ('iNKT', 'MAIT', 'Vd1', and 'Vd2'), and subtypes of the 'Empty Cells' group, the 'Other Tissue' group, the 'Dead Cells' group, and the 'Mixed Population' group.

Our results revealed that when trained with sufficient reference datasets, the ANN-SCT-PBMC model is robust and can survive a small number of non-representative instances hidden in the training set. The model can discriminate between and assess the relative representativeness of SCT data when it has only been trained on high-quality reference datasets. The confounders of different properties can have varying effects on model vulnerability. The factors that can affect model vulnerability include - the proportion of reference and non-representative datasets, the proportion of the classes in training and testing sets, the similarity of gene expression between cell types and subtypes, and the properties of non-representative datasets, etc.

7.2 Introduction

The quality and representativeness of data has an impact on the classification performance of supervised machine learning artificial neural network (ANN) models [268, 269]. In the process of studying using ANN for PBMC classification based on SCT gene expression profiles (ANN-SCT-PBMC classification), we found that non-representative data (cells with confounding factors such as 'empty cells', 'other tissue', 'dead cells', 'activated cells', and 'mixed population') can be easily mixed in the data set. It can have implications for accurate classification of PBMC using ANN models at single-cell resolution. The presence of non-representative data can affect model training and prediction results.

This study attempts to explore the relationship between the vulnerability and robustness of the ANN-SCT-PBMC model and the representativeness of the datasets. Meanwhile, this study designed four parallel external cross-validation experiments to investigate the specific effects of non-representative components on model performance when they were included in SCT datasets from different sources.

This study aims to identify:

1. The effect of non-representative data to ANN-SCT-PBMC classification performance: the model performance in four parallel external cross-validation experiments (4-supersets-swapping) when progressively eliminating non-representative data of different properties.

2. The specific factors affecting the vulnerability of the ANN-SCT-PBMC model.

3. With the gradual elimination of non-representative datasets, the robustness of the ANN-SCT-PBMC model for the five classes (BC, DC, MC, NK, and TC) in the 4-supersets-swapping experiment.

4. With the gradual elimination of non-representative datasets, the robustness of the ANN-SCT-PBMC model for different cell subtypes in the 4-supersets-swapping experiment.

7.3 Materials and Methods

This study focuses on the vulnerability testing and robustness validation of ANN model, with the effect of different groups of non-representative PBMC SCT data sets. This study is an extension

of previous studies [16, 65]. In this study, the entire data sets have included five groups of non-representative data sets and one group of 56 clean reference data sets (the same as the healthy PBMC samples used in incremental learning study [146]).

The fundamental architecture of ANN classifier and the assessment metrics of classification performance are the same as in earlier research.

The 56 clean data sets [146] have 10x Demo, GEO, BroadS1, and BroadS2, four data sources. The five groups of non-representative data sets represent groups of “Empty Cells”, “Other Tissue”, “Dead Cells”, “Activated Cells”, and “Mixed Population”, sourcing from GEO database. These groups contain common PBMC datasets that are easily confused and misused as reference datasets. In this study, they were used to test the influence of the representativeness of the training set and the confounding factors on the classification model.

The study design has involved comparative vulnerability testing using both non-representative data sets and clean data sets, with the method of four supersets swapping [146].

7.3.1 Study design

We deployed a "from noise to clean" experimental design to validate and examine the vulnerability and robustness of ANN-SCT-PBMC classification model.

In the first round of the experiment, the datasets for training and testing consist of all clean datasets and non-representative datasets. All datasets are divided into four super sets according to the data source, and four parallel ANN training and testing steps (Steps 1-4, Figure 48 B) are performed in 4-super-sets-swapping manner – three super sets are used as training set, while use the fourth super set to test the trained network, then iteratively swap the next super set as testing set. After one round of 4-super-sets-swapping training and testing, it collects the classification results to each testing set, and evaluates model performance of this round. Then enter the second round. In this round, one non-representative data set is eliminated from all datasets, and the ANN training and testing of 4-super-sets-swapping is performed again. The same steps are then repeated, cumulatively removing the next non-representative data set in the next round until the final round - only clean reference datasets exist. The detailed workflow is shown in Figure 48 A). The order of decreasing deletion of the non-representative data sets is based on arbitrary order, from the least representative to the closest to clean data, in the order of eliminating: ‘Empty Cells’ → ‘Other Tissue’ → ‘Dead Cells’ → ‘Activated Cells’ → ‘Mixed Population’ (as shown in Figure 49).

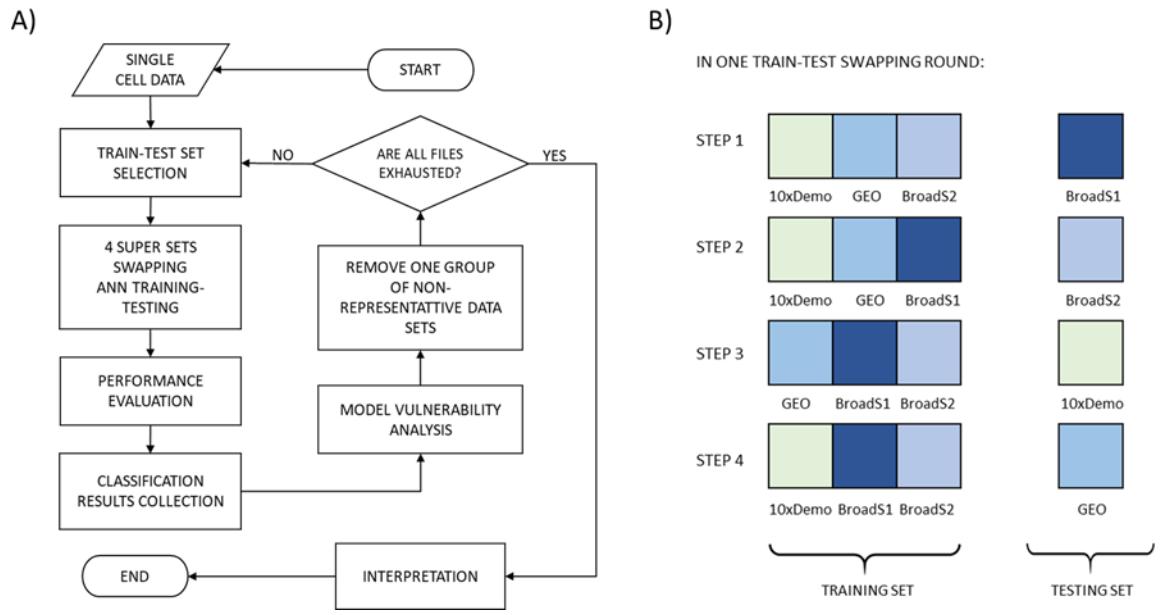


Figure 48. Schematic diagram of study design. A) shows the workflow of model training and testing. Classification results are collected and analyzed with various trained neural networks and testing sets in different rounds. B) demonstrates the components of training set and testing set in one round of four-super-sets-swapping. There are four steps in one round. As an example, in Step 1, the sum of 10xDemo, GEO, and BroadS2 are used as training set, while BroadS1 is used as testing set to assess the classification accuracy.

In this study, there are in total 17 rounds of 4-super-sets-swapping ANN training and testing. As an example, 'Round 1' (as shown in Figure 49) is the first round of ANN training and testing, in the first step of it (Step 1, Figure 48 B): 9 reference data sets (of 10x data source); 11 reference data sets, 50 empty cells, and GSM3162632 [270], GSM3162630 [270], GSM3087629 [49], GSM3430548 [256], GSM3430549 [256], GSM3478792 [255], GSM3558027 [255], GSM3258345 [257], GSM3258347 [257], GSM3258346 [257], GSM3258348 [257], and GSM3087628 [49] (of GEO data source); 31 reference data sets (of BroadS2 data source); these (as ticked with check marks in Figure 49) are used to train the network. The complete BroadS1 data sets are used as the testing set.

In Step 2 (Figure 48 B), 31 reference data sets (of BroadS2) are used as testing set, others are used as training set. Similarly, in Step 3 and 4 (Figure 48 B), data sets of 10xDemo and of GEO, are used to test their corresponding trained networks, respectively. In the following Round 2 to Round 17 (Figure 49), the eliminated data in each round (each column in the figure) is illustrated as blank (Figure 49). The non-representative data is eliminated one at a time in the rounds.

The last round (Round 17) includes 4-super-set-swapping train-test on 56 clean reference data sets. The seventeen rounds of 4-super-sets-swapping train-test experiments were done until all non-representative data sets were eliminated. The voting results of the trained neural networks were collected and analyzed of each step in each round.

SOURCE	INDEX	DATA SETS	PROPERTY	ROUND-1	ROUND-2	ROUND-3	ROUND-4	ROUND-5	ROUND-6	ROUND-7	ROUND-8	ROUND-9	ROUND-10	ROUND-11	ROUND-12	ROUND-13	ROUND-14	ROUND-15	ROUND-16	ROUND-17
10x	SRP073767	9-Data-Sets	Clean	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
BroadS1	SCP345	5-Data-Sets		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
BroadS2	SCP424/5/6	31-Data-Sets		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
GEO	GEO	11-Data-Sets		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
GEO	N/A	25-Empty-Cells	Empty Cells	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	N/A	15-Empty-Cells		✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	N/A	5-Empty-Cells		✓	✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	N/A	5-Empty-Cells		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	GSM3162632	Tumor_Ascites_DC	Other Tissue	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	GSM3162630	Tonsil_DC		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	GSM3087629	Methanol_SSC_T8	Dead Cells	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	GSM3430548	Donor1_ IL-10-Producing_Foxp3-_T4	Activated Cells	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	GSM3430549	Donor2_ IL-10-Producing_Foxp3-_T4		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	GSM3478792	Nonmalignant_P5_CD3+CD5intSSCint_T4		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	GSM3558027	Nonmalignant_P5_CD3+CD5intSSCint_T4_Afth		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	GSM3258345	HLA-DR	Mixed Population	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	GSM3258347	HLA-DR_Control		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	GSM3258346	CD19		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	GSM3258348	CD19_Control		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	GSM3087628	CD8		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Figure 49. Illustration of involved data sets of ANN train-test in Round 1 to 17. In the final round of 4-super-sets-swapping, solely 56 reference data sets were included. The study design aims on testing the vulnerability of ANN-SCT-PBMC classification model with confounding factors on data representativeness.

The results of Round 5, 7, 8, 12, 17 (in Figure 49) represents the model performance when iteratively cumulatively depleting 'Empty Cells, 'Other Tissue', 'Dead Cells', 'Activated Cells', and 'Mixed Population' data groups. For these rounds, we also used 1-Sensitivity [271, 272] to assess the classification error rate of each cell subtype:

$$1 - Sensitivity = 1 - \frac{TP}{TP + FN}$$

where, TP – true positives (class positives classified as positives), FN – false negatives (class positives but predicted as negatives).

For Round 1 to 17, we performed assessment with confusion matrix, accuracy (ACC), specificity, sensitivity/RE, PR, and F1-score, same as in previous studies [65, 146]. Specifically, we used accuracy (ACC) for multi-class overall assessment and used F1-score for individual cell type assessment (i.e., for BC, DC, MC, NK, and TC).

The comparative analysis within Round 1 to 17 demonstrated the vulnerability and robustness of ANN classifier with the effect of SCT PBMC data representativeness.

7.3.2 Data

The 56 clean data sets [146] representing PBMC from healthy blood samples were selected. Their data set group property is described as "clean" in this study.

The other 17 data sets are considered as "non-representative" data sets, they are sourced from GEO database and form "Empty Cells", "Other Tissue", "Dead Cells", "Activated Cells", and "Mixed Population" five non-representative data groups. The datasets were collected and standardized to 30,698 gene list, and converted to five different file formats, in this study, MTX file format was mainly used for ANN training and testing, considering computational efficiency. The gene expression of each cell profile used in training and testing is filtered and standardized raw gene counts (quality control), captured and sequence aligned by 10x SCT protocol.

For "Empty Cells", we put 10, 5, 2, and 1 empty cells under each class (BC, DC, MC, NK, and TC) of GEO data, in the Round 1, 2, 3, and 4, individually. The four rounds contained 50, 25, 10, and 5 empty cells in total, respectively. From the Round 5, 'Empty Cells' noise is not included in

the loop. These round-reduced empty cells were populated with the labels of five classes and treated as five non-representative datasets. The gene expression of the empty cells is zero, simulating the effect of "dropout" instances (in real-life single cell sequencing situation) to the ANN SCT classification model.

Two dendritic cells data sets have formed "Other Tissue" group, one is 'tumor ascites dendritic cells (GSM3162632) [270]' and the other is 'tonsil dendritic cells (GSM3162630) [270]'. They are tissue-residential dendritic cells samples, the SCT gene expression of those dendritic cells are different from those of peripheral blood circulating dendritic cells.

The data set GSM3087629 [49] represents "Dead Cells", the biological sample of it is CD8+ T cells of healthy frozen PBMCs fixed with methanol reagent. After processing with methanol fixation, the cells are pictured with specific instantaneous gene expression status, that is different from the gene expression level of fresh cells or frozen-thawed cells.

GSM3430548 [256], GSM3430549 [256], GSM3478792 [255], and GSM3558027 [255] represent for "Activated Cells" data group. GSM3430548 and GSM3430549 are IL-10 producing Foxp3-CD4+ T cells from healthy blood samples of two donors, they are specifically selected activated CD4+ T cells for functional study. GSM3478792 and GSM3558027 are nonmalignant P5 CD3+CD5intSSCintCD4+ T cells from fresh blood of a 61-year-old male patient donor, pre- and post- stage IVA Sézary syndrome (T4N1M0B2) treatment. The CD4+ T cells in those two data sets are in activated functional status, their gene expression can be different from normal circulating CD4+ T cells in healthy individual samples.

In "Mixed Population" group, there are five data sets - GSM3258345 [257], GSM3258347 [257], GSM3258346 [257], GSM3258348 [257], and GSM3087628 [49]. The first four data sets come from one series GSE116683 [257]. GSM3258345 and GSM3258347 are pair data sets of HLA-DR+ cells, GSM3258347 is the control group. They are designed to target live enriched HLA-DR+ cells and deplete other blood lineages (CD235a, CD3, CD4, CD8, CD19, CD56). They are mixed populations of cells expressed HLA-DR cell surface receptor. Monocytes constitutively express HLA-DR, those two data sets are labeled under "MC" class. GSM3258346 and GSM3258348 are pair data sets of CD19+ cells, they are enriched and selected by FACS cell sorting, that solely targeting live CD19+ cells and depleting other blood lineages (CD235a, CD3, CD4, CD8, HLA-DR, CD56). They are labeled with "BC" class, as CD19 is typical cell protein marker of B cells. They are mixture of various cell populations expressed CD19 protein marker, other than B cells expressed CD19 marker. Those four data sets are sampled from healthy fresh blood. GSM3087628 is a mixture of cell groups expressed CD8 protein marker, that is sorted by magnetic beads of MACS cell sorting. It is labeled as the "TC" class, as CD8 is a regular marker

of T cells.

The total number of cells in this study is 145,605. Table 16 summarized the data sets and cells numbers of each class involved in this study. Table 17 shows as a brief metadata for 17 non-representative reference data sets, it includes information such as series ID, publication date, cell type and the labeling class.

Table 16. An overview of the 73 SCT data sets used in this study is as below. Cell numbers and the number of data sets are shown for each class.

NUMBER OF CELLS AND DATA SETS OF CLASSES						
Sources	**BC**	**DC**	**MC**	**NK**	**TC**	**Total Cells**
10x Demo	10,085 (1)	0	2,612 (1)	8,385 (1)	64,347 (6)	85,429 (9)
GEO	1,796 (3)	4,362 (3)	3,311 (4)	319 (2)	24,912 (15)	34,700 (27)
BroadS1	1,660 (1)	142 (1)	1,661 (1)	1,394 (1)	8,326 (1)	13,183 (5)
BroadS2	1,884 (4)	271 (8)	2,132 (8)	842 (4)	7,164 (8)	12,293 (32)
Total	15,425 (9)	4,775 (12)	9,716 (14)	10,940 (8)	104,749 (30)	145,605 (73)

Table 17. The summary of the 17 non-representative data sets.

SOURCE	SERIES	DATE	CELL TYPE	CLASS
GEO	N/A	N/A	10-Empty-Cells-in-BC	BC
			10-Empty-Cells-in-DC	DC
			10-Empty-Cells-in-MC	MC
			10-Empty-Cells-in-NK	NK
			10-Empty-Cells-in-TC	TC
	GSM3162632	5/30/2018	Tumor Ascites Dendritic Cells	DC
	GSM3162630		Tonsil Dendritic Cells	
	GSM3087629	7/25/2018	CD8+ T Cells (Methanol SSC)	TC
	GSM3430548	11/7/2018	IL-10 Producing Foxp3-CD4+ T Cells (Donor 1)	TC
	GSM3430549		IL-10 Producing Foxp3-CD4+ T Cells (Donor 2)	
	GSM3478792	1/31/2019	Nonmalignant P5 CD3+CD5intSSCintCD4+ T Cells	TC
	GSM3558027	7/25/2019	Nonmalignant P5 CD3+CD5intSSCintCD4+ T Cells (After Therapy)	
	GSM3258345	10/15/2018	HLA-DR+ Cells	MC
	GSM3258347		HLA-DR+ Cells (Control)	
	GSM3258346		CD19+ Cells	BC
	GSM3258348		CD19+ Cells (Control)	
	GSM3087628	7/25/2018	CD8+ Cells	TC

The Figure 50 shows the cell subtypes and their proportions in four data sources. In clean data sets, there are 4 subtypes (‘BC’ of 10x, ‘Bn’/ ’Bm’ of BroadS1, ‘BC’ of BroadS2) of B cells, 3 subtypes (‘DC’ of BroadS1, ‘DC’/ ’pDC’ of BroadS2) of dendritic cells, 6 subtypes (‘M14’ of 10x, ‘M14’ of GEO, ‘M14’/ ’M16’ of BroadS1, ‘M14’/’M16’ of BroadS2) of monocytes, 4 subtypes (‘NK’ of 10x, ‘NK’ of GEO, ‘NK’ of BroadS1, ‘NK’ of BroadS2) of NK cells, and 24 subtypes (‘CD45RA+CD25-T4naive’/ ‘T4’/ ‘CD45RA+T8naive’/ ‘T8’/ ‘CD45RO+T4mem’/ ‘CD4+CD25+Treg’ of 10x, ‘T4’/ ‘T8’/ ‘iNKT’/ ‘MAIT’/ ‘Vd1’/ ‘Vd2’/ ‘T4’/ ‘CCR5+CD69-T4’ of GEO, ‘aTreg’/ ‘nonT’/ ‘rTreg’/ ‘T4em’/ ‘T4naive’/ ‘T8em’/ ‘T8naive’/ ‘Tncl’ of BroadS1, and ‘T4’/ ‘T8’ of BroadS2) of T cells.

In 17 experimental data sets (highlighted in yellow in Figure 50), it has other 3 subtypes of dendritic cells, 7 of T cells, 3 of monocytes, and 3 of B cells. The hierarchical relationship of these cell subtypes has been drawn in the ontology of PBMC [146].

In the four super sets (10x, GEO, BroadS1, and BroadS2), the frequency of cell numbers in each class (BC, DC, MC, NK, and TC) are corresponded to the reference values of healthy interval ranges in PBMC, as described in previous studies [65, 146].

DATA SOURCE	CELL SUBTYPE	SUBTYPE NUMBER	CLASS	FREQUENCY	TOTAL NUMBER
10x (CLEAN)	BC	10085	BC	11.81%	85423
	M14	2612	MC	3.06%	
	NK	8385	NK	9.82%	
	CD45RA+CD25-T4naive	10479	TC	75.32%	
	T4	11213			
	CD45RA+T8naive	11953			
	T8	10209			
	CD45RO+T4mem	10224			
	CD4+CD25+Treg	10263			
GEO (ALL)	M14	856	MC	2.47%	34700
	NK	309	NK	0.89%	
	T4	222	TC	9.01%	
	T8	310			
	iNKT	325			
	MAIT	382			
	Vd1	284			
	Vd2	204			
	T4	965			
	CCR5+CD69-T4	435			
	10-Empty-Cells-in-BC	10	BC	0.03%	
	10-Empty-Cells-in-DC	10	DC	0.03%	
	10-Empty-Cells-in-MC	10	MC	0.03%	
	10-Empty-Cells-in-NK	10	NK	0.03%	
	10-Empty-Cells-in-TC	10	TC	0.03%	
	Tumor_Ascites_DC	1613	DC	12.54%	
	Tonsil_DC	2739			
	Methanol_SSC_T8	4753	TC	46.44%	
	Donor1_ IL-10-Producing_Foxp3-_T4	1247			
	Donor2_ IL-10-Producing_Foxp3-_T4	1902			
	Nonmalignant_P5_CD3+CD5intSSCint_T4	4486			
	Nonmalignant_P5_CD3+CD5intSSCint_T4_Afth	3725			
	HLA-DR	48	MC	7.05%	
	HLA-DR_Control	2397			
	CD19	26	BC	5.15%	
	CD19_Control	1760			
	CD8	5662	TC	16.32%	
BroadS1 (CLEAN)	Bn	1169	BC	12.59%	13183
	Bm	491			
	DC	142	DC	1.08%	
	M14	1263	MC	12.60%	
	M16	398			
	NK	1394	NK	10.57%	
	aTreg	921	TC	63.16%	
	nonT	426			
	rTreg	1072			
	T4em	975			
	T4naive	1134			
	T8em	1031			
	T8naive	1336			
	Tncl	1431			
BroadS2 (CLEAN)	BC	1884	BC	15.33%	12292
	DC	202	DC	2.20%	
	pDC	68			
	M14	1809	MC	17.34%	
	M16	323			
	NK	842	NK	6.85%	
	T4	3380	TC	58.28%	
	T8	3784			

Figure 50. The cell subtypes and proportions in each data source. Subtypes of one same class are highlighted in similar color hue. The color bar shows the level of abundance in 'Subtype Number', 'Frequency', and 'Total Number'.

In four super sets swapping, the testing sets can have 85,423 cells (10x), 34,700 cells (GEO), 13,183 cells (BroadS1) or 12,292 cells (BroadS2). The training sets can have 133,306 cells ({10x U GEO U BroadS1}), 132,415 cells ({10x U GEO U BroadS2}), 110,898 cells ({10x U BroadS1 U BroadS2}), 60,175 cells ({GEO U BroadS1 U BroadS2}).

7.4 Results

7.4.1 Overall accuracy of four testing sets in each round

The results of overall ANN classification are shown in Figure 51. It shows the prediction accuracy of the testing set for four parallel train-test steps, within seventeen swapping rounds.

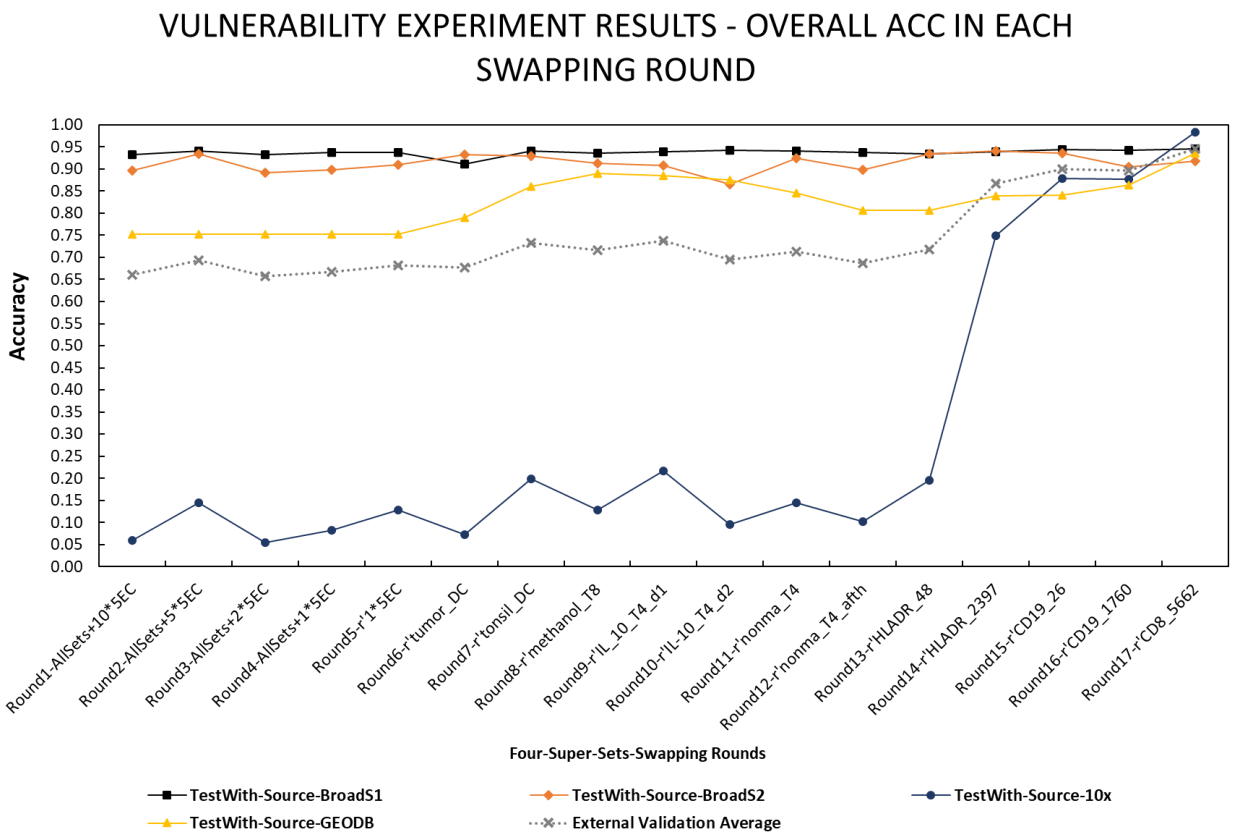


Figure 51. Accuracy of 4-super-sets-swapping in Round 1 to 17. The predication on BroadS1 and BroadS2 testing sets showed high accuracy (averagely 0.937 and 0.914 in seventeen rounds). With the representativeness of data sets increased during seventeen rounds, the model performance on 10x testing set had significant improvement, from 0.059 in Round 1 to 0.983 in Round 17. The average of the external validation to four sets showed upward trend on overall accuracy. All four data sets showed a trend of convergence, eventually reaching over 0.917. In the final round, the average accuracy of the four supersets reached 0.945.

With cumulatively eliminating non-representative data sets in training set, the classification accuracy of **testing set BroadS1** (the black line in Figure 51) remained above 0.912 across seventeen rounds. The average prediction accuracy of BroadS1 testing set was 0.937 for PBMC 5-class classification.

The classification performance on **BroadS2** data sets overall remained above 0.866. With the adjustment and alteration in the training set, the prediction results for the BroadS2 data sets fluctuated, but the overall classification performance remained relatively high, with an average

accuracy of 0.914 in total seventeen rounds.

The prediction performance on **the 10x Demo testing set** has shown a significant improvement across the seventeen rounds, the overall accuracy has increased from 0.059 in Round 1 to 0.983 in Round 17 (Figure 51). From Round 1 to 5, by removing 'Empty Cells' in the training set, the prediction performance on 10x improved by 0.069 of accuracy. Considering the large data proportion of clean data sets in the training set (60,125 reference cells of 60,175 total cells, 99.92%, as shown in Figure 49 and Figure 50), the ANN model was sensitive and vulnerable to 'Empty Cells' confounding factor hidden in the training data. When testing with 10x data sets, the model vulnerability was largely affected by representativeness of the training data. During Round 1 to Round 12, with the groups 'Empty Cells, 'Other Tissue', 'Dead Cells', and 'Activated Cells' included in the training set, overall accuracy on 10x Demo data sets swinged up and down around 0.119. Different numbers of empty cells and different noise properties of the non-representative instances in the training set have irregular negative effects on classification accuracy. Since R12, there was a rapid increase in accuracy, until the R17 accuracy rose to 0.983. From R12 to R17, the training set gradually removed the data sets of 'Mixed Population' group, one at a time.

For GEO testing set, the neural networks in seventeen rounds were trained by the reference data sets of 10x, BroadS1, and BroadS2 (as shown in Figure 49). The entire classification results on GEO testing set showed an overall upward trend. From Round 1 to 17, it increased 24.41% of accuracy, when eliminating confounding data sets in both training and testing sets, within 4-super-sets swapping experiments. The results of GEO in the seventeen rounds demonstrated the effect of the components of testing set to model accuracy evaluation in multi-class classification.

The gray line in Figure 51 showed **the average** accuracy of the 4-super-sets-swapping external validation results. During Round 1 to Round 17, it demonstrated a steadily increase in overall accuracy. With the improvement of data representativeness, the overall accuracy rose from 0.660 to 0.945, for four independent super sets train-test swapping experiments.

From Figure 51, the performance for four testing sets all converged to above 0.917 at the last swapping round. Taken together, when with high data representativeness (solely included clean reference data sets), the external validation accuracy of four independent sets for ANN-SCT-PBMC 5-class classification ranged from 0.917 to 0.983, with the average of 0.945.

7.4.2 F1-score of individual cell types in each round

We measured F1-score value of each cell type (BC, DC, MC, NK, and TC) prediction for seventeen swapping rounds. F1-score is the harmonic mean of precision and recall, in our 5-class classification, it was the main metric used in individual cell type evaluation. The results of F1-score of each class in each round for four parallel testing have shown as Figures 52-55.

7.4.2.1 Testing with BroadS1

When the training set included data source of 10x, GEO, and BroadS2, testing with BroadS1 (Figure 52), the prediction performance of BC, MC, TC was quite robust, F1-score steadily remained 0.943 to 0.983, averagely 0.961. The F1-score of NK class was around 0.773, for seventeen rounds.

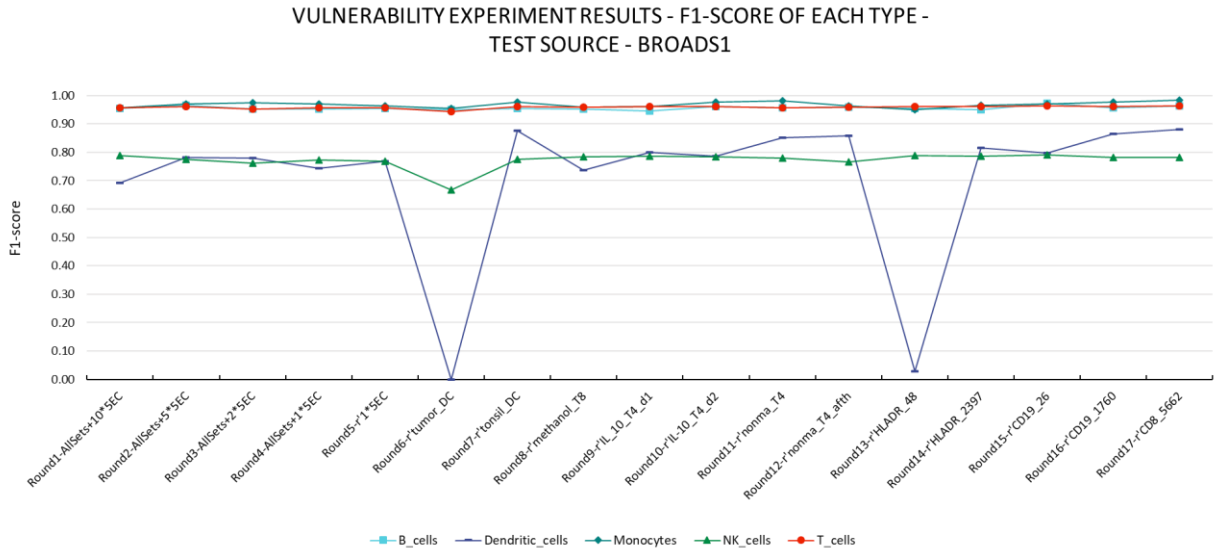


Figure 52. F1-score results of five cell types in 4-super-sets-swapping rounds, with BroadS1 as the testing set. The prediction performance of BC, MC, NK, and TC were stable, while it of DC was close to zero in Round 6 and 13. The F1-score of BC, MC, TC were kept around 0.961, and it of NK was around 0.773, during seventeen rounds.

The classification to 142 dendritic cells of BroadS1 was affected by non-representative data sets in the training set. It was unstable, it was 0.000 of F1-score measure in Round 6 and 0.027 in Round 13, while remaining 0.693 to 0.880 for other rounds. When gradually removed 30,408 of non-reference cells out of 132,415 of total cells (22.96%), the model classification performance was not affected much, when it comes to BC, MC, NK, and TC.

The DC prediction was fragile, while gradually removed 4,362 of non-reference dendritic cells out of 4,632 total dendritic cells in training set. With a small amount of instances, the model behavior on DC was quite vulnerable and it was largely affected by the number, proportion, and properties of the non-reference data of five classes, that were hidden in the training set.

7.4.2.2 Testing with BroadS2

When we used BroadS2 as the testing set and the data sourcing from 10x, GEO, and BroadS1 as the training set, the prediction results (Figure 53) on each cell type was quite similar to the experiments when testing with BroadS1 (Figure 52). From Figure 53, the F1-score on BC, MC, and TC during seventeen rounds stabilized around 0.947, compared to 0.961 when tested with BroadS1 (Figure 52). The F1-score to NK demonstrated a slightly more up-and-down trend – averagely 0.681, with the lowest value of 0.536 in Round 10.

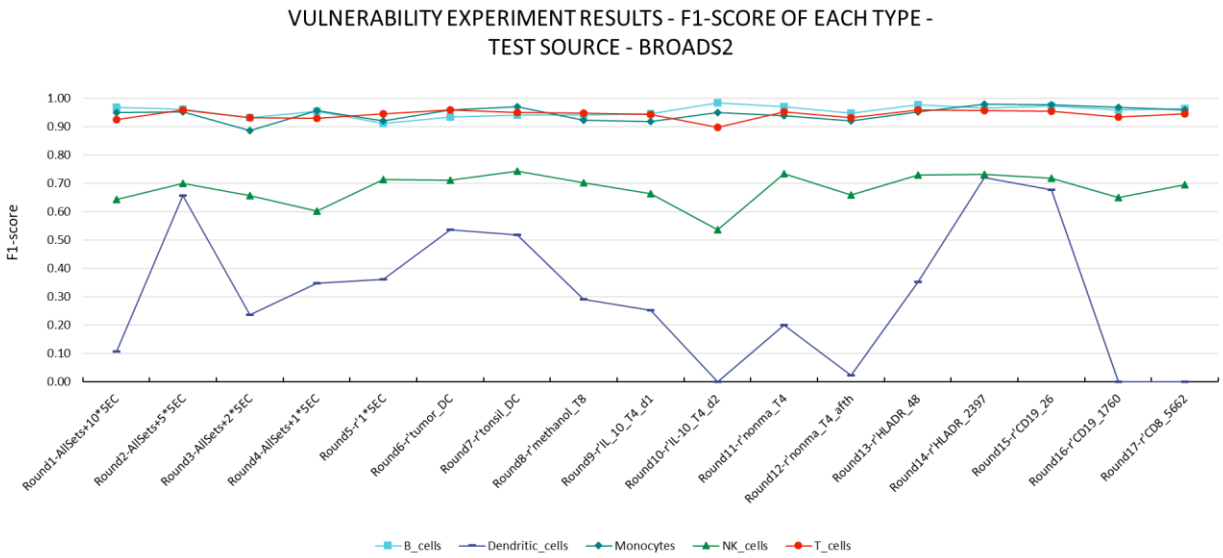


Figure 53. F1-score of five cell types in 4-super-sets-swapping rounds, with BroadS2 as the testing set. The classification performance of BC, MC, and TC class were stable, it remained around 0.947. The F1-score of NK class was around 0.681, during seventeen swapping rounds. The model prediction of DC was irregular, that was 0.310 in average.

In Round 10, both the F1-score of NK and TC decreased, the NK F1-score decreased by 0.127, the TC F1-score decreased by 0.045, compared to Round 9. In Round 11, the F1-score of NK and TC prediction increased back to 0.734 and 0.952, respectively. The ANN model was sensitive to changes in the representativeness of the gene expression profiles that comprise the training set.

In Round 10, the training set included 4,486 cells of the data set 'Nonmalignant_P5_CD3+CD5intSSCint_T4' and 3,725 cells of 'Nonmalignant_P5_CD3+CD5intSSCint_T4_Afth' (both of the group 'Activated Cells'), under TC class. The existing of the set 'Nonmalignant_P5_CD3+CD5intSSCint_T4' confounded the model pattern recognition ability on NK-TC binary classification.

This set was a T cell set while sampled from patient fresh blood – a 61-year-old male patient donor, with stage IVA Sézary syndrome (T4N1M0B2) being treated. The gene expression of this T cell set was different from it of normal healthy T cell set, that caused the misclassification between NK and TC – as in Round 10, 793 more T cells in BroadS2 (that has totally 7,164 T cells) was predicted as NK cells, compared to Round 9.

In Round 11, the training set eliminated the data set 'Nonmalignant_P5_CD3+CD5intSSCint_T4' while kept the data set 'Nonmalignant_P5_CD3+CD5intSSCint_T4_Afth', that was the pair T cell set of the patient after therapy. The gene profile of patient T cell set after therapy demonstrated less influence on model vulnerability. The inclusion of 3,725 after-therapy T cells increased the classification performance of NK and TC.

The prediction on DC class showed irregular results, the F1-score of DC was 0.310 in average, during seventeen swapping rounds. Similar to when testing with BroadS1, the DC prediction was largely affected by the non-representative data of five classes, in the training set.

7.4.2.3 Testing with 10x

The results of 10x testing set showed as Figure 54, that had F1-score results of four classes – BC, MC, NK, and TC. All four classes showed a trend from a low initial F1-score value (averagely 0.036) to a gradual increase until it converged to a high F1-score value (averagely 0.948). The results of 10x testing set clearly showed the significant impact of non-representative data sets to ANN-SCT-PBMC classification model – when gradually purifying and cleaning training set from non-reference data, the classification ability for each class was improved, and it reached the highest point when there were only clean reference sets included in the training set (as shown in Figure 54, in Round 17, the F1-score for BC, MC, NK, and TC classification, was 0.969, 0.873, 0.954, and 0.995, respectively).

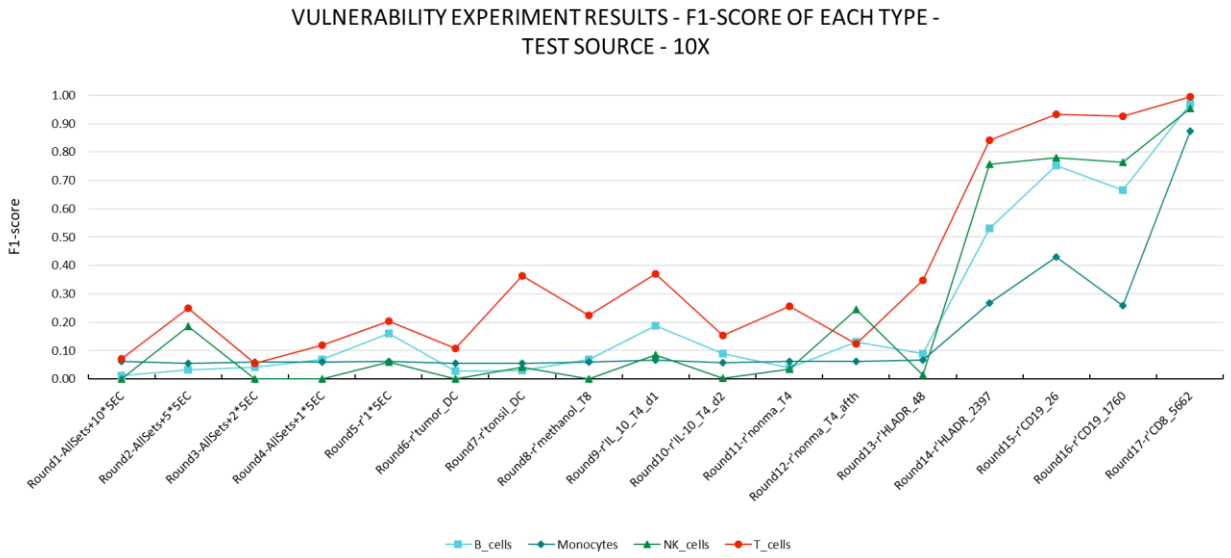


Figure 54. F1-score of four cell types in 4-super-sets-swapping rounds, with 10x as the testing set. The results showed the impact of groups of non-representative data on ANN-SCT-PBMC classifier, especially when it accounts for a large proportion of the training set.

As listed in Figure 50, 10x data source contains 85,423 cells, which accounts for a large proportion in the data composition of four sources (58.67% of the sum of all data sets). The 85,423 cells of 10x set are qualified reference gene profiles. When the 10x set was not included in training set (Figure 54), ANN-SCT-PBMC model was heavily impacted by the proportion of reference data sets in training set – that was 49.47% in Round 1, while 100.00% in Round 17. Unlike when the large reference set 10x was included in the training set and maintained basic robustness for BC, MC, NK, and TC prediction (Figure 52 and Figure 53), the model was vulnerable in 10x testing experiments (Figure 54) – that trained with the combination of GEO, BroadS1, and BroadS2.

In Round 16, without the balancing benefits from other classes, when solely the 5662 cells of data set 'CD8' (of 'Mixed Population' group) included in non-reference sets, the model was affected largely – the F1-score for all four classes was decreased, by 0.086, 0.172, 0.015, and 0.007, for BC, MC, NK, and TC, individually. In Round 16, the model was trained by 13,183 cells of BroadS1, 12,292 cells of BroadS2, 4,292 reference cells of GEO, and 5662 CD8 cells of GEO. The 'CD8' cells are the mixture of sorted cell populations that expressed CD8 protein marker. The CD8 receptor exists on the surface of different cell types within PBMC, including NK cells, innate-like T cells, cytotoxic CD8+ T cells, dendritic cells [273], and that caused the confusion on prediction to BC, MC, NK and TC classes.

7.4.2.4 Testing with GEO

The predictions of the four classes showed a gradual convergence trend from Round 5, and reached the maximum value in the last round (Round 17, Figure 55).

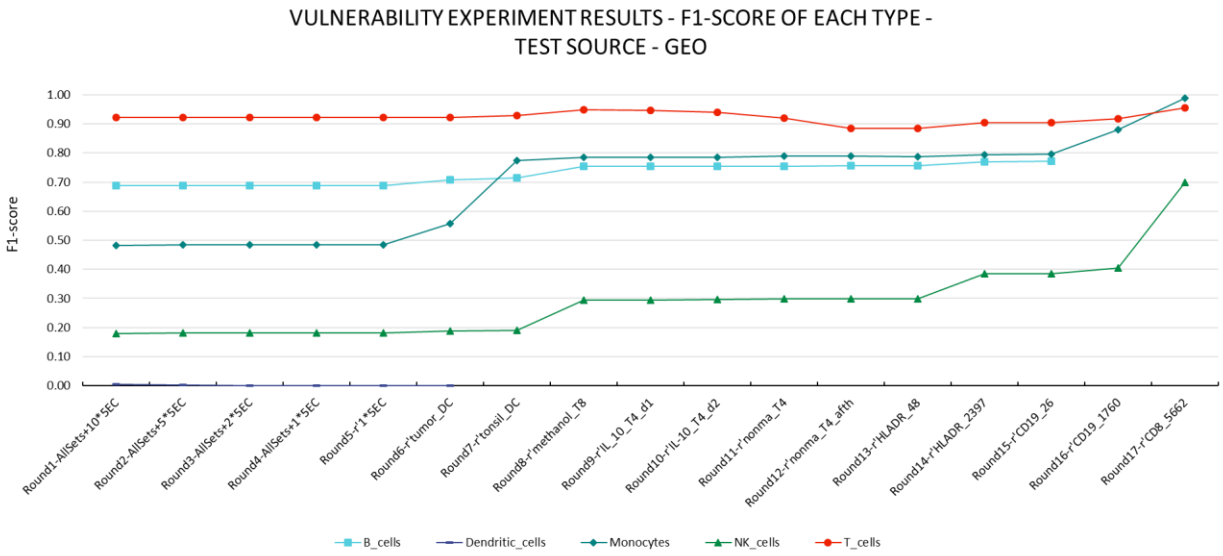


Figure 55. F1-score of five cell types in 4-super-sets-swapping rounds, with GEO as the testing set. The model demonstrated pattern recognition ability in distinguishing representative and non-representative data in GEO, after being jointly trained by 10x, BroadS1, and BroadS2 reference data sets. The F1-score of four classes (BC, MC, NK, and TC) showed a trend of increasing and convergence within seventeen rounds. In Round 17, the F1-score of MC, NK, and TC reached 0.989, 0.700, and 0.955.

As shown in Figure 55, the model jointly trained by the 10x, BroadS1, and BroadS2 reference data sets had certain pattern recognition ability for the representative data and non-representative data in GEO. The model had good classification performance on representative data in GEO, while had low performance on non-representative data. In Round 17, after gradually eliminating non-representative sets of five groups, the F1-score value for MC, NK, and TC was 0.989, 0.700, and 0.955, respectively. Generally, the F1-score of BC and MC kept around 0.702~0.729, the F1-score of TC remained averagely around 0.922, and it of NK class steadily increased from 0.180 in Round 1 to 0.700 in Round 17. The prediction F1-score of DC class kept around 0.001, as shown in Figure 55. The 1,613 cells of 'Tumor_Ascites_DC' data set and the 2,739 cells of 'Tonsil_DC' data set were correctly not predicted as DC class, that demonstrated the pattern recognition ability of the

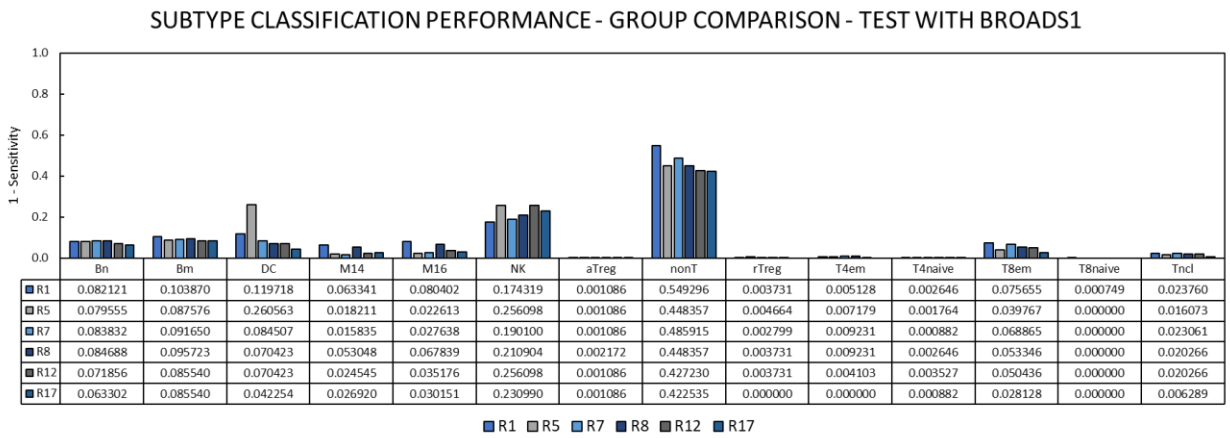
model. The SCT gene expression profiles of ‘Tumor_Ascites_DC’ and ‘Tonsil_DC’ data sets are different from those of healthy circulating dendritic cells of PBMC. These two sets are dendritic cells sampled from tumor ascites and tonsil tissue. Additionally, the calculation result of F1-score was also affected by the imbalance of multi-class classification.

7.4.3 Subtype classification performance in Round 1, 5, 7, 8, 12, and 17 – group comparison

The classification evaluation to each cell subtype was measured by *1-Sensitivity*, that is used as measurement for error rate. We measured the value of *1-Sensitivity* of subtypes in Round 1, 5, 7, 8, 12, and 17, specifically. These are the rounds when each entire group of non-representative sets was eliminated. For example, in Round 12, the entire group of ‘Activated Cells’ was removed from 4-super-sets-swapping train-test experiment, as compared to Round 8, that included ‘Activated Cells’ and ‘Mixed Population’ groups. Group comparisons of subtype error rates in these rounds demonstrated the robustness of the model to different subtypes when faced with changes in data profiles across groups.

7.4.3.1 Subtype performance of testing set BroadS1

There are 14 cell subtypes in the testing set BroadS1, as shown in Figure 56. Within the group comparison of Round 1, 5, 7, 8, 12, and 17, the subtype error rate (refers to *1-Sensitivity* in the study) showed an overall downward trend – i.e., the model performance for subtypes generally improved as the non-representative groups were pulled out. Among them, the subtypes ‘NK’ and ‘nonT’ (Figure 56) had high error rate across the six rounds, with an average of 0.220 and 0.464, respectively.



	Bn	Bm	DC	M14	M16	NK	aTreg	nonT	rTreg	T4em	T4naive	T8em	T8naive	Tncl
R1	0.082121	0.103870	0.119718	0.063341	0.080402	0.174319	0.001086	0.549296	0.003731	0.005128	0.002646	0.075655	0.000749	0.023760
R5	0.079555	0.087576	0.260563	0.018211	0.022613	0.256098	0.001086	0.448357	0.004664	0.007179	0.001764	0.039767	0.000000	0.016073
R7	0.083832	0.091650	0.084507	0.015835	0.027638	0.190100	0.001086	0.485915	0.002799	0.009231	0.000882	0.068865	0.000000	0.023061
R8	0.084688	0.095723	0.070423	0.053048	0.067839	0.210904	0.002172	0.448357	0.003731	0.009231	0.002646	0.053346	0.000000	0.020266
R12	0.071856	0.085540	0.070423	0.024545	0.035176	0.256098	0.001086	0.427230	0.003731	0.004103	0.003527	0.050436	0.000000	0.020266
R17	0.063302	0.085540	0.042254	0.026920	0.030151	0.230990	0.001086	0.422535	0.000000	0.000000	0.000882	0.028128	0.000000	0.006289

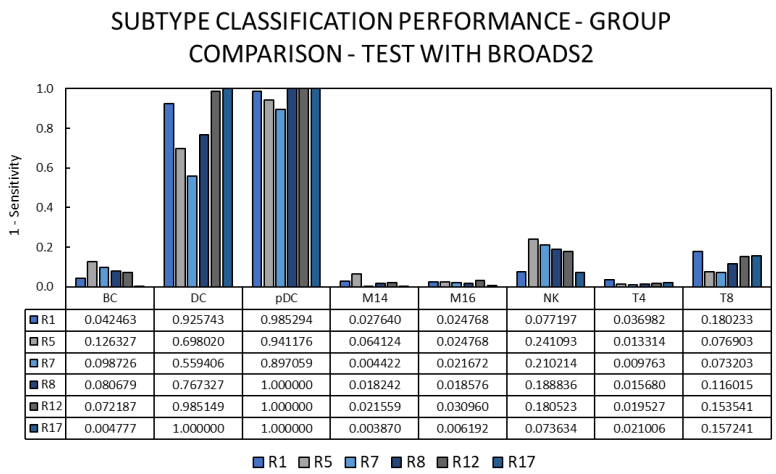
Figure 56. The performance of subtype prediction within group comparisons, used BroadS1 as testing set. The subtypes 'NK' and 'nonT' had high error rate, 0.220 and 0.464 in average. ANN model steadily recognized subtype patterns, with various non-representative sets included in training set, in six rounds.

Even in Round 17, 99.07% of the misclassifications in the NK class was 'T cells', which is related to the biological similarity hidden in the gene expression profiles of NK cells and T cells. As expected, 'nonT' had a high classification error rate, roughly half of 'nonT' were classified as 'NK cells' and the other half were classified as 'T cells', in all six rounds. There was a potential paradox in original annotation of 'nonT' subtype: there were two labelling methods for the BroadS1 dataset, one of which annotates the 'nonT' cell population as 'non-T cells', while the other method identifies them as 'T cells'. This group of cells has specific gene expression intermediate between NK cells and T cells.

Taken together, the results for BroadS1 subtypes indicated that the model can sensitively identify cell populations with confounding gene expression profiles, to a certain extent. Furthermore, the model showed robustness across group comparisons in six rounds.

7.4.3.2 Subtype performance of testing set BroadS2

The classification results for the 8 subtypes of BroadS2 varied widely in six rounds. The subtypes 'DC' and 'pDC' had extremely high error rates (the average over six rounds were 0.823 and 0.971). The 'BC', 'NK', and 'T8' exhibited average error rate as 0.071, 0.162, and 0.126, respectively.



	BC	DC	pDC	M14	M16	NK	T4	T8
R1	0.042463	0.925743	0.985294	0.027640	0.024768	0.077197	0.036982	0.180233
R5	0.126327	0.698020	0.941176	0.064124	0.024768	0.241093	0.013314	0.076903
R7	0.098726	0.559406	0.897059	0.004422	0.021672	0.210214	0.009763	0.073203
R8	0.080679	0.767327	1.000000	0.018242	0.018576	0.188836	0.015680	0.116015
R12	0.072187	0.985149	1.000000	0.021559	0.030960	0.180523	0.019527	0.153541
R17	0.004777	1.000000	1.000000	0.003870	0.006192	0.073634	0.021006	0.157241

Figure 57. The performance of subtype prediction within group comparisons, taken BroadS2 as testing set.

Compared to Round 1 and Round 5, the error rates of ‘DC’ and ‘pDC’ were significantly decreased in Round 7, which excluded the ‘Empty Cells’ data group, and ‘Tumor_Ascites_DC’ and ‘Tonsil_DC’ data sets of the ‘Other Tissue’ group. The ‘empty-cells’, ‘non-healthy’, and ‘non-peripheral’ sets had a greater impact on the prediction of DC than those confounding factors of other groups. At the same time, due to the small sample size (‘sample’ refers to data samples), the DC class was more affected by non-representative datasets, showing larger vulnerability in the six rounds.

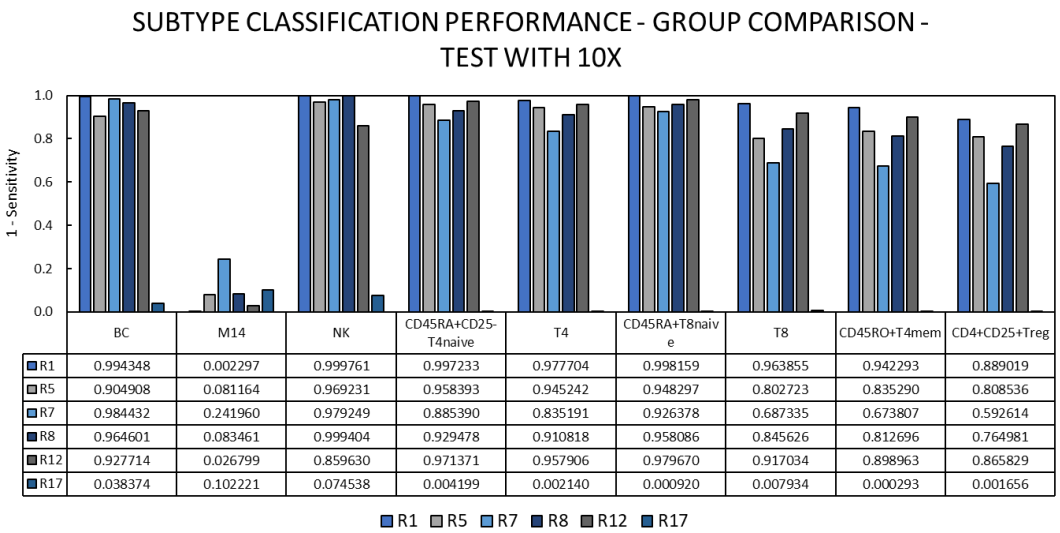
As the number of samples of non-representative T cells gradually decreased, the predictions of subtypes ‘NK’ and ‘T8’ exhibited a “trade-off” trend - the ‘NK’ error rate decreased, while the ‘T8’ prediction error rate increased.

In general, when BroadS2 was used as the testing set, the vulnerability of ANN model was affected by the number of samples within the category, the type of non-representative data, and the similarity of gene expression profiles.

7.4.3.3 Subtype performance of testing set 10x

With groups of confounding factors included, the 9 subtypes of the 10x testing set had high error rate in Round 1, 5, 7, 8, and 12. Among these rounds, the average error rate of ‘BC’, ‘NK’, and 6 T cell subtypes was 0.902. While in Round 17, the subtype error rate of the 10x testing set showed

a sharp drop, with an average of 0.026 for the 9 subtypes. When the 10x dataset (that has a large sample size) was not included in the training set, the model robustness was significantly affected by the non-representative data sets (in Round 1, 5, 7, 8, and 12).



	BC	M14	NK	CD45RA+CD25-T4naive	T4	CD45RA+T8naive	T8	CD45RO+T4mem	CD4+CD25+Treg
R1	0.994348	0.002297	0.999761	0.997233	0.977704	0.998159	0.963855	0.942293	0.889019
R5	0.904908	0.081164	0.969231	0.958393	0.945242	0.948297	0.802723	0.835290	0.808536
R7	0.984432	0.241960	0.979249	0.885390	0.835191	0.926378	0.687335	0.673807	0.592614
R8	0.964601	0.083461	0.999404	0.929478	0.910818	0.958086	0.845626	0.812696	0.764981
R12	0.927714	0.026799	0.859630	0.971371	0.957906	0.979670	0.917034	0.898963	0.865829
R17	0.038374	0.102221	0.074538	0.004199	0.002140	0.000920	0.007934	0.000293	0.001656

Figure 58. The performance of subtype prediction within group comparisons, used 10x as testing set. The average error rate of 9 subtypes decreased from 0.863 to 0.026, when gradually excluded non-representative sets from experiments. The variations in the types and proportions of non-representative datasets had a significant impact on the model's robustness.

The error rate of 6 T cell subtypes (‘CD45RA+CD25-T4naive’, ‘T4’, ‘CD45RA+T8naive’, ‘T8’, ‘CD45RO+T4mem’, and ‘CD4+CD25+Treg’, Figure 58) dropped in Round 7 (average value 0.767) and then rose again in Round 8 and Round 12 (average value 0.870 and 0.932, respectively). The robustness of the model was affected heavily by variations in the types and proportions of non-representative datasets within the five classes.

7.4.3.4 Subtype performance of testing set GEO

When testing with GEO data source, the five non-representative groups were included in the testing set. The network was trained with clean reference data sets of 10x, BroadS1, and BroadS2.

In the testing set, there are 11 subtypes from reference datasets and 13 subtypes from non-representative datasets.

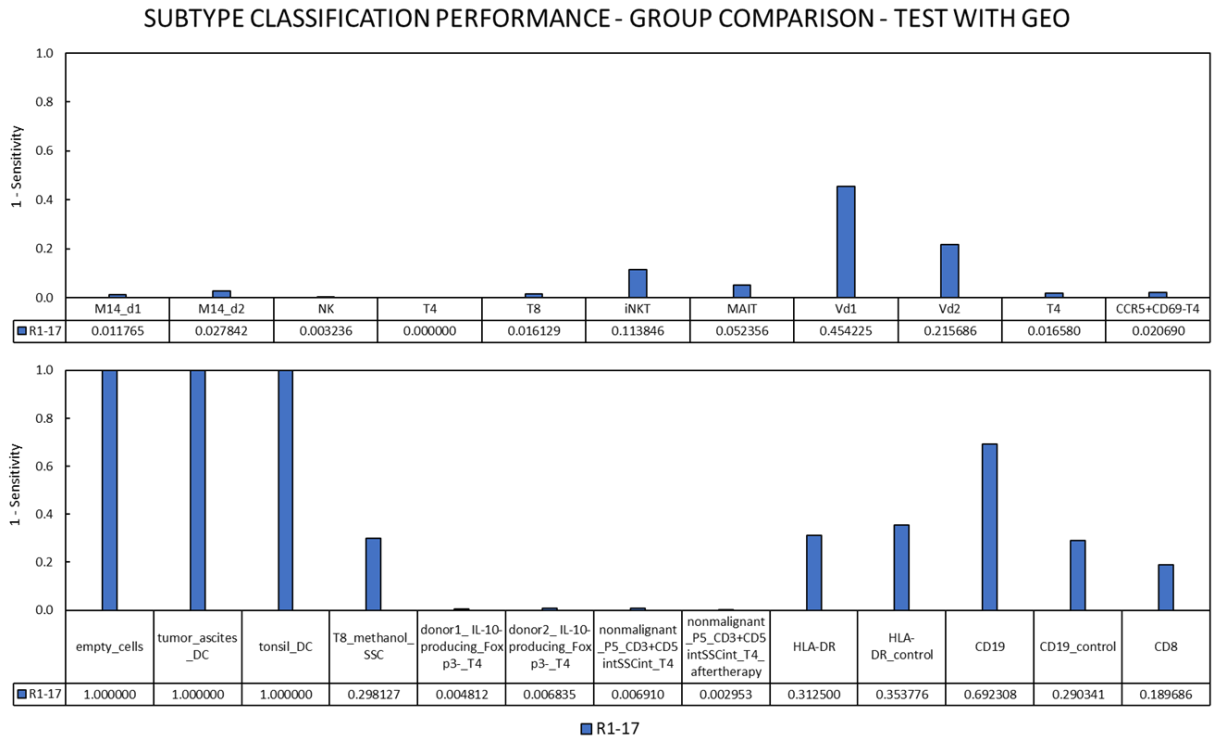


Figure 59. The performance of subtype prediction within group comparison, testing with GEO.

In the reference set subtypes, the misclassification was concentrated in the four innate-like T cell subtypes - 'iNKT', 'MAIT', 'Vd1', and 'Vd2' (the average of *1-Sensitivity* value was 0.209). They have special gene expressions different from those of conventional T cells.

Among the non-representative subtypes, misclassification occurred mainly in the 'Empty Cells' group, the 'Other Tissue' group, the 'Dead Cells' group, and the 'Mixed Population' group (Figure 59). For these groups, the values of *1-Sensitivity* were 1.000, 1.000, 0.298, and 0.368, individually.

The results indicated that, ANN model trained on high-quality reference datasets have a certain ability to screen and identify the representativeness of SCT data. The voting results of the neural network trained with high-quality instances can be used to evaluate the SCT data representativeness.

7.5 Conclusions

7.5.1 Overall accuracy

Overall, the results indicated that the non-representativeness of data can negatively affect the ANN-SCT-PBMC model classification performance. The model was vulnerable and had low classification accuracy when there were non-representative instances included in the datasets (overall average accuracy was 0.660 in Round 1, Figure 51). As the non-representative data was gradually stripped from the datasets, the average accuracy gradually increased, across the four external cross validation experiments, eventually converging to 0.945 (Figure 51).

When high-quality reference data accounts for more than half of the total training instances (e.g. 10x dataset, accounting for 58.67% of the sum of all datasets), the model is robust against changes in attributes and proportions of non-representative components hidden in the training set. As from the results, the five-class classification average accuracy of BroadS1, BroadS2, and GEO testing set fluctuated between 0.912~0.946, 0.866~0.941, 0.752~0.935, respectively; while the fluctuation range of 10x testing set was relatively large, between 0.054~0.983 (Figure 51).

7.5.2 F1-score of 5 classes

From the F1-score of each cell type, while being affected by non-representative instances, the class with small scale (the “rare class”) is more vulnerable (e.g. the DC class had irregular and unstable predictions, Figure 52 and Figure 53). The performance of model for rare class can be greatly influenced by the attributes and proportions of the data.

When the training set contains reference data source with large cardinality, the model is robust to the predictions of BC, MC, NK, and TC classes and remains stable over 17 rounds (Figure 52, Figure 53, and Figure 55). Compared with BC, MC and TC classes, NK prediction had lower F1-score results. Due to the similar SCT gene expression profile to TC instances, the prediction performance of NK was greatly restricted.

When the training set contains large number of non-representative instances, with the continuous reduction of non-representative instances and the increase of high-quality reference instances, the F1-score for BC, MC, NK and TC demonstrated a gradual increase and convergence in 17 rounds, with a final average of 0.948 (Figure 54).

7.5.3 Performance on subtypes

From the results of the six group comparisons, it can be seen that the classification performance for subtypes varies with the properties and proportions of different non-representative groups.

In the BroadS1 testing set, subtype misclassification occurred mainly in 'NK' and 'nonT', that was traced to the highly confounding gene expressions of NK cells and T cells.

Meanwhile, the 'DC' and 'pDC' subtypes in BroadS2 consistently had high error rate across 17 train-test rounds, with an average of 0.823 and 0.971, respectively (Figure 57). Compared to other non-representative data groups, the 'Empty Cells' group and the non-representative DC instances in the 'Other Tissue' group had a greater impact on DC class prediction. The error rate for the two DC subtypes both decreased when these groups were excluded from the training set.

When the large reference set 10x was excluded from the training set, the non-representativeness of the dataset has a significant effect on model performance. The 9 subtypes of the 10x testing set had high error rates in Round 1, 5, 7, 8, and 12. In Round 17, the subtype error rate dropped dramatically, with an average of 0.026 for the 9 subtypes (Figure 58).

From the results of GEO subtypes, it can be clearly seen that the model trained by high-quality reference datasets has a certain ability to identify and evaluate the representativeness of SCT data. The model had low error rates for subtypes of the reference datasets and high error rates for non-representative datasets. Misclassifications focused on four innate-like T cell subtypes 'iNKT', 'MAIT', 'Vd1', and 'Vd2'; one subtype of 'Empty Cells' group; two subtypes of 'Other Tissue' group; one subtype of 'Dead Cells' group; and five subtypes of 'Mixed Population' group. It indicated that the model was more vulnerable to non-representative instances from 'Empty Cells', 'Other Tissue', 'Dead Cells', and 'Mixed Population' groups, than the 'Activated Cells' group.

7.5.4 Final overall conclusions

Comprehensively, the ANN-SCT-PBMC model is robust when trained with sufficient reference instances, it can tolerate a small number of non-representative instances hidden in the training set. Among the five classes, the prediction performance of the rare class can fluctuate greatly. At the same time, the model purely trained by high-quality reference sets has the ability to distinguish and evaluate the relative representativeness of SCT data. Of the five confounding factors, the 'Empty Cells', 'Other Tissue', 'Dead Cells', and 'Mixed Population' groups can have greater

influence than the ‘Activated Cells’ group.

In final conclusion, in this study, the factors that can affect the vulnerability of the ANN-SCT-PBMC model include

a. the proportion of the reference datasets and the non-representative datasets in the training set,

b. the proportion of the classes in the training set and the testing set,

c. the similarity of gene expression between cell types and cell subtypes,

d. the properties of the non-representative datasets (the least relevant non-representative datasets can have a higher impact and the specific impact needs to be confirmed by further study).

7.6 Discussion

This study demonstrates the effect of decreasing non-representative datasets one by one on the robustness of the ANN-PBMC-SCT model in four external cross-validation experiments. The results found that the ratio of reference and non-representative datasets has a large impact on model performance. As shown in Figure 60, when the reference datasets occupy a large proportion of the training set, the model can counteract the negative effects of non-representative instances (Figure 60, A and B); while when the non-representative datasets occupy a large proportion of the training set, the model's vulnerability increases with the number of non-representative instances (Figure 60, C). More in-depth discussions can include – investigating the number of reference instances required to train a qualified ANN-SCT-PBMC model, and the number of non-representative instances it can tolerate.

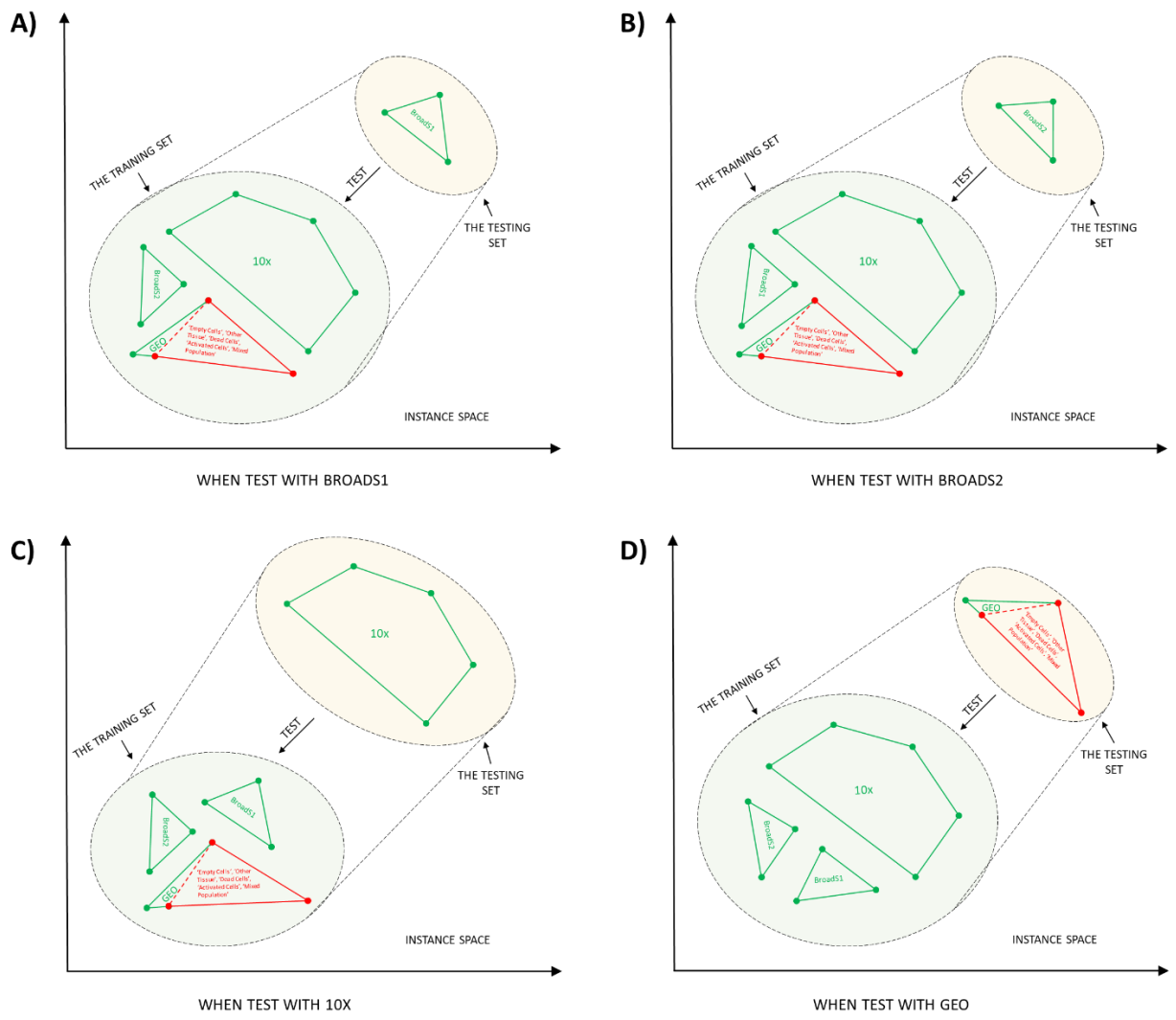


Figure 60. The illustration for the effect of the proportion of reference and non-representative datasets on model performance. The A), B), C), and D) represent the specifics of the training and testing sets in four external cross-validation experiments in this study when non-representative instances are involved. Different symbol sizes imply the relative proportions of different data sources (e.g., 10x, BroadS1, BroadS2, GEO reference set, and GEO non-representative set account for roughly 59%, 9%, 8%, 3%, and 21% of total instances).

Meanwhile, the classification results on the GEO testing set indicate that the model trained on sufficient pure reference data has the ability to evaluate the representativeness of SCT data (Figure 60, D). The voting results of the model can be used as a metric for scoring the representativeness of the dataset [269].

A limitation of this study is that the experimental design shows only one potential order of cumulative reduction of the five groups of confounders, and results under other alternative orders can be done in further studies - our focus of this study is to reveal the trends in the performance changes brought about by the accumulation of non-representative datasets. The confounders of different properties have different effects on model vulnerability. The impact of individual confounders on model performance can be explored in further study.

It is worth noting that the non-representative datasets used in our study only represents part of the SCT samples, and more instances from other sources are needed to complete further validation with larger sample size.

Furthermore, in addition to the five confounding factors included in this study (the 'Empty Cells', 'Other Tissue', 'Dead Cells', 'Activated Cells', and 'Mixed Population' groups), model performance is also affected by other factors (described in SCT cell ontology), such as

a. the "Maturation status: Immature/Transitional/Mature" in "Cell Properties" dimension;

b. the "Developmental stage: Fetal/Pediatric/Young/Middle-age/Elderly" in "Organism Properties" dimension;

c. or the "Sample preparation: Isolation/Staining-and-purity-assessment/Cell-sorting" in "Experimental Settings" dimension; etc.

The effect of these other confounding factors on the ANN-SCT-PBMC model vulnerability needs to be explored further.

CHAPTER 8 GENERAL CONCLUSIONS AND FUTURE WORK

8.1 General Conclusions

This research demonstrated and proved the concept that single cell classification can be done with purely supervised ML method ANN and multi-source independent SCT data. The ANN-SCT-PBMC classification models have achieved good performance with various datasets generated from multisource studies. It has demonstrated adequate gene expression profile pattern recognition and classification ability, also good robustness to SCT datasets with diverse sample conditions.

This research collected and standardized PBMC SCT **reference datasets** from various data sources (GEO database, Broad Institute, and 10x Genomics Demonstration), with five main cell types (B cells, dendritic cells, monocytes, natural killer (NK) cells, and T cells). Corresponding **metadata** has been organized for the qualitative description and statistical properties of SCT datasets. We designed and described the multi-dimensional single-cell **ontology** for PBMC SCT classification. It used over 163 dimensions to category and characterize single cells, based on prior knowledge in immunology and single cell domain. **In the pilot study,** we used 27 SCT datasets of 121,281 single cell instances to achieve the accuracy of classification of PBMC of 89.4% and proved the concept that using purely supervised machine learning method to classify single cells. **In the initial study of incremental learning,** we selected 27 SCT datasets that derived from healthy PBMC samples. We used methods of cyclical holdout internal cross-validation, external validation, and validation on added datasets to evaluate SCT classification performance. The cyclical incremental learning that simulating real-life situation by the gradual addition of new independent data sets to ANN training improved classification. In the final cycle, the overall accuracy reached 93.0% for 4-class classification. **In the follow-up expanded incremental learning study,** we sorted solely clean representative data and newly collected dataset BroadS2 and explored the effect of different data processing protocols to ANN models. BroadS2 dataset has brought reference dendritic cells into the training sets. With 56 clean reference datasets and seven cycles of training and testing, the overall accuracy of 5-class classification reached 94.6%. Classification accuracy for B cells, monocytes, and T cells exceeded 95%. Classification accuracy of NK cells kept around 75% caused by the similarity between NK cells and T cell subsets. The accuracy of dendritic cells was limited due to small proportion of numbers in the training sets. We also analyzed the impact of different processing methods to gene expression profiles and SCT classification. The results indicated that datasets derived from minimally processed samples (PBMC separation only) contributed to SCT gene expression pattern recognition. **Building upon these,** we used other 17 non-representative datasets of five groups: ‘empty cells’, ‘other tissue’,

‘dead cells’, ‘activated cells’, and ‘mixed population’, and 17 rounds of four parallel external cross-validation (four-supersets-swapping) experiments to explore the **vulnerability** of ANN-SCT-PBMC classification models. Our findings showed that the ANN-SCT-PBMC model was robust and could tolerate non-representative instances hidden in the training set when trained with sufficient reference datasets. When the model has been trained on purified high-quality reference data, it can distinguish and evaluate the representativeness of SCT data. The factors that affected model vulnerability include - the proportion of reference and non-representative datasets, the proportion of the classes in training and testing sets, the similarity of gene expression between cell types and subtypes, and the properties of non-representative datasets, etc.

Overall, our research demonstrates that supervised ML ANN is a viable option for single cell classification. This research gives solution to the current “eleven grand challenges” of SCT data analysis. It built reference datasets for PBMC SCT classification. It solves the difficulties in single cell classification using purely supervised ML ANN, that demonstrates generalization and robustness on various upcoming data sets.

Cell ontology and biological explanation with gene expression profile were used to comprehend the performance of ANN classifier. We found that other than the ‘cell properties’ (inherent gene expression of cell types), other dimensions in cell ontology can have significant impact on SCT classification performance, such as - data generation protocol (cell sorting), tissue source (peripheral circulating or tissue-residential), cell state (healthy, methanol fixation, or functionally activated), cell labeling (mixed population).

The results revealed that well-defined, rigorous, and detailed annotation of true classes is the key issue of ANN SCT classification. The results indicated that adequate reference data, produced under exacting and stringent SCT protocols, and labeled with a comprehensive and in-depth multi-dimensional cell ontology are necessary for highly accurate single cell classification, which can support future predictive health development. The machine-simulated purely supervised single cell classification models can maximize the potential value of SCT data, it can help achieve future systematic regular detection of human health, early disease diagnosis and prevention, as well as development in hematology.

8.2 Future Work

Our work has limitations as start-up research in the field, further study could be done on:

1. Data: Need more reference data sets. With more SCT data sets of multi-dimensional subtypes of PBMC, a classification model based on multi-dimensional PBMC cell ontology can be built and evaluated with metrics.

2. Model: This study proves the concept of using SCT data and ANN to do supervised single cell classification. Optimized methods with model structure and parameter changing or comparison with different supervised ML methods can be used to explore the performance of SCT classification.

3. Metadata: This study focuses on healthy PBMC SCT data training and testing, focusing on proof-of-concept validation and generating benchmark reference data for data quality control and disease/function PBMC data pattern recognition. When it comes to potential further functional study situations, the model can be trained with disease data sets (sample of CLL patients), and used for disease single cell prediction.

4. Incremental learning: In this study, we deployed the traditional incremental learning – manual data accumulation. We aimed on observing model performance on independent SCT datasets. Combined reference data on specific dimension of cell ontology, ensemble learning can be used in research on model learning efficiency.

5. Class imbalance: In this study, we kept data class distribution as collected, simulating the real frequency of each cell type in human blood. A study on balanced class classification can be explored with under-sampling, over-sampling, and advanced-sampling methods.

6. Divide and conquer: Further explore the misclassification of TC and NK, MC and BC, and the identification and differentiation of intermediate cell subtypes.

7. Model vulnerability: Further explore the effect of other dimensions (in the multi-dimensional cell ontology) on SCT classification performance, such as ‘maturation status’, ‘developmental stage’, ‘gender’, etc.

REFERENCES

[1] D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz, L. Pinello, P. Skums, A. Stamatakis, C. S.-O. Attolini, S. Aparicio, J. Baaijens, M. Balvert, B. d. Barbanson, A. Cappuccio, G. Corleone, B. E. Dutilh, M. Florescu, V. Guryev, R. Holmer, K. Jahn, T. J. Lobo, E. M. Keizer, I. Khatri, S. M. Kielbasa, J. O. Korbel, A. M. Kozlov, T.-H. Kuo, B. P. F. Lelieveldt, I. I. Mandoiu, J. C. Marioni, T. Marschall, F. Mölder, A. Niknejad, L. Raczkowski, M. Reinders, J. d. Ridder, A.-E. Saliba, A. Somarakis, O. Stegle, F. J. Theis, H. Yang, A. Zelikovsky, A. C. McHardy, B. J. Raphael, S. P. Shah, and A. Schönhuth, "Eleven grand challenges in single-cell data science," *Genome Biology,* vol. 21, no. 1, pp. 31, 2020/02/07, 2020.

[2] V. Y. Kiselev, T. S. Andrews, and M. Hemberg, "Challenges in unsupervised clustering of single-cell RNA-seq data," *Nature Reviews Genetics,* vol. 20, no. 5, pp. 273-282, 2019/05/01, 2019.

[3] A. Kulkarni, A. G. Anderson, D. P. Merullo, and G. Konopka, "Beyond bulk: a review of single cell transcriptomics methodologies and applications," *Current Opinion in Biotechnology,* vol. 58, pp. 129-136, 2019/08/01/, 2019.

[4] M. D. Luecken, and F. J. Theis, "Current best practices in single-cell RNA-seq analysis: a tutorial," *Molecular systems biology,* vol. 15, no. 6, pp. e8746, 2019.

[5] V. Svensson, R. Vento-Tormo, and S. A. Teichmann, "Exponential scaling of single-cell RNA-seq in the past decade," *Nature Protocols,* vol. 13, no. 4, pp. 599-604, 2018/04/01, 2018.

[6] A. Peng, X. Mao, J. Zhong, S. Fan, and Y. Hu, "Single-Cell Multi-Omics and Its Prospective Application in Cancer Biology," *Proteomics,* vol. 20, no. 13, pp. 1900271, 2020.

[7] N. Navin, and J. Hicks, "Future medical applications of single-cell sequencing in cancer," *Genome medicine,* vol. 3, no. 5, pp. 1-12, 2011.

[8] X. Tang, Y. Huang, J. Lei, H. Luo, and X. Zhu, "The single-cell sequencing: new developments and medical applications," *Cell & Bioscience,* vol. 9, no. 1, pp. 53, 2019/06/26, 2019.

[9] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, and A. Siddiqui, "mRNA-Seq whole-transcriptome analysis of a single cell," *Nature methods,* vol. 6, no. 5, pp. 377-382, 2009.

[10] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, and J. Zhu, "Massively parallel digital transcriptional profiling of single cells," *Nature communications,* vol. 8, no. 1, pp. 1-12, 2017.

[11] S. A. Morris, "The evolving concept of cell identity in the single cell era," *Development,* vol. 146, no. 12, pp. dev169748, 2019.

[12] M. Efremova, R. Vento-Tormo, J.-E. Park, S. A. Teichmann, and K. R. James, "Immunology in the era of single-cell technologies," *Annual review of immunology,* vol. 38, pp. 727-757, 2020.

[13] Y. Ando, A. T.-J. Kwon, and J. W. Shin, "An era of single-cell genomics consortia," *Experimental & Molecular Medicine,* vol. 52, no. 9, pp. 1409-1418, 2020.

[14] V. Svensson, R. Vento-Tormo, and S. A. Teichmann, “Exponential scaling of single-cell RNA-seq in the past decade,” *Nat Protoc,* vol. 13, no. 4, pp. 599-604, Apr, 2018.
[15] P. See, J. Lum, J. Chen, and F. Ginhoux, “A single-cell sequencing guide for immunologists,” *Frontiers in immunology,* vol. 9, pp. 2425, 2018.
[16] R. A. Shaikh, J. Zhong, M. Lyu, S. Lin, D. Keskin, G. Zhang, L. Chitkushev, and V. Brusic, "Classification of Five Cell Types from PBMC Samples using Single Cell Transcriptomics and Artificial Neural Networks." pp. 2207-2213.
[17] X. Wang, Y. He, Q. Zhang, X. Ren, and Z. Zhang, “Direct comparative analyses of 10X Genomics Chromium and smart-seq2,” *Genomics, Proteomics & Bioinformatics*, 2021.
[18] J. Ding, X. Adiconis, S. K. Simmons, M. S. Kowalczyk, C. C. Hession, N. D. Marjanovic, T. K. Hughes, M. H. Wadsworth, T. Burks, and L. T. Nguyen, “Systematic comparison of single-cell and single-nucleus RNA-sequencing methods,” *Nature biotechnology,* vol. 38, no. 6, pp. 737-746, 2020.
[19] P. Angerer, L. Simon, S. Tritschler, F. A. Wolf, D. Fischer, and F. J. Theis, “Single cells make big data: New challenges and opportunities in transcriptomics,” *Current Opinion in Systems Biology,* vol. 4, pp. 85-91, 2017.
[20] W. Wang, and G.-Z. Wang, “Understanding molecular mechanisms of the brain through transcriptomics,” *Frontiers in physiology,* vol. 10, pp. 214, 2019.
[21] D. T. Paik, S. Cho, L. Tian, H. Y. Chang, and J. C. Wu, “Single-cell RNA sequencing in cardiovascular development, disease and medicine,” *Nature Reviews Cardiology,* vol. 17, no. 8, pp. 457-473, 2020.
[22] D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, and A. Mahfouz, “Eleven grand challenges in single-cell data science,” *Genome biology,* vol. 21, no. 1, pp. 1-35, 2020.
[23] D. Osumi-Sutherland, C. Xu, M. Keays, A. P. Levine, P. V. Kharchenko, A. Regev, E. Lein, and S. A. Teichmann, “Cell type ontologies of the Human Cell Atlas,” *Nature Cell Biology,* vol. 23, no. 11, pp. 1129-1135, 2021.
[24] M. Mosallaei, N. Ehtesham, S. Rahimirad, M. Saghi, N. Vatandoost, and S. Khosravi, “PBMCs: A new source of diagnostic and prognostic biomarkers,” *Archives of physiology and biochemistry*, pp. 1-7, 2020.
[25] C. R. Kleiveland, "Peripheral blood mononuclear cells," *The impact of food bioactives on health*, pp. 161-167: Springer, Cham, 2015.
[26] R. Petryszak, M. Keays, Y. A. Tang, N. A. Fonseca, E. Barrera, T. Burdett, A. Füllgrabe, A. M.-P. Fuentes, S. Jupp, and S. Koskinen, “Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants,” *Nucleic acids research,* vol. 44, no. D1, pp. D746-D752, 2016.
[27] S. Melzer, S. Zachariae, J. Bocsi, C. Engel, M. Löffler, and A. Tárnok, “Reference intervals for leukocyte subsets in adults: results from a population-based study using 10-color flow cytometry,” *Cytometry Part B: Clinical Cytometry,* vol. 88, no. 4, pp. 270-281, 2015.
[28] L. Yang, Y. Zhang, N. Mitic, D. B. Keskin, G. L. Zhang, L. Chitkushev, and V. Brusic, "Single-cell mRNA Profiles in PBMC." pp. 1318-1323.
[29] L. Minjie, M. Radenkovic, D. B. Keskin, and V. Brusic, "Classification of single cell types during leukemia therapy using artificial neural networks." pp. 1258-1261.
[30] B. Zheng, L. Minjie, L. Sen, and V. Brusic, "Tissue of origin classification from single cell mRNA expression by Artificial Neural Networks." pp. 1346-1350.

[31] G. Nardo, S. Pozzi, M. Pignataro, E. Lauranzano, G. Spano, S. Garbelli, S. Mantovani, K. Marinou, L. Papetti, and M. Monteforte, "Amyotrophic lateral sclerosis multiprotein biomarkers in peripheral blood mononuclear cells," *PloS one,* vol. 6, no. 10, pp. e25545, 2011.

[32] D. A. Giles, M. E. Moreno-Fernandez, T. E. Stankiewicz, S. Graspeuntner, M. Cappelletti, D. Wu, R. Mukherjee, C. C. Chan, M. J. Lawson, and J. Klarquist, "Thermoneutral housing exacerbates nonalcoholic fatty liver disease in mice and allows for sex-independent disease modeling," *Nature medicine,* vol. 23, no. 7, pp. 829-838, 2017.

[33] F. Porichis, M. G. Hart, M. Griesbeck, H. L. Everett, M. Hassan, A. E. Baxter, M. Lindqvist, S. M. Miller, D. Z. Soghoian, and D. G. Kavanagh, "High-throughput detection of miRNAs and gene-specific mRNA at the single-cell level by flow cytometry," *Nature communications,* vol. 5, no. 1, pp. 1-12, 2014.

[34] J. Pourahmad, and A. Salimi, "Isolated human peripheral blood mononuclear cell (PBMC), a cost effective tool for predicting immunosuppressive effects of drugs and xenobiotics," *Iranian journal of pharmaceutical research: IJPR,* vol. 14, no. 4, pp. 979, 2015.

[35] P. Sen, E. Kemppainen, and M. Orešič, "Perspectives on systems modeling of human peripheral blood mononuclear cells," *Frontiers in molecular biosciences,* vol. 4, pp. 96, 2018.

[36] P. J. Delves, S. J. Martin, D. R. Burton, and I. M. Roitt, *Roitt's essential immunology*: John Wiley & Sons, 2017.

[37] F. Betsou, A. Gaignaux, W. Ammerlaan, P. J. Norris, and M. Stone, "Biospecimen Science of Blood for Peripheral Blood Mononuclear Cell (PBMC) Functional Applications," *Current Pathobiology Reports,* vol. 7, no. 2, pp. 17-27, 2019/06/01, 2019.

[38] B. Mesko, S. Poliska, and L. Nagy, "Gene expression profiles in peripheral blood for the diagnosis of autoimmune diseases," *Trends in molecular medicine,* vol. 17, no. 4, pp. 223-233, 2011.

[39] J. C. Rockett, M. E. Burczynski, A. J. Fornace Jr, P. C. Herrmann, S. A. Krawetz, and D. J. Dix, "Surrogate tissue analysis: monitoring toxicant exposure and health status of inaccessible tissues through the analysis of accessible tissues and cells," *Toxicology and applied pharmacology,* vol. 194, no. 2, pp. 189-199, 2004.

[40] X. Lin, H. Yu, C. Zhao, Y. Qian, D. Hong, K. Huang, J. Mo, A. Qin, X. Fang, and S. Fan, "The peripheral blood mononuclear cell count is associated with bone health in elderly men: A cross-sectional population-based study," *Medicine,* vol. 95, no. 15, 2016.

[41] J. Xu, R. A. Shaikh, and V. Brusic, "Single Cell Transcriptomics Reveals Summary Patterns Specific for PBMCs and Other Cell Types." pp. 1435-1438.

[42] A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, and M. Clatworthy, "Science forum: the human cell atlas," *elife,* vol. 6, pp. e27041, 2017.

[43] M. J. T. Stubbington, O. Rozenblatt-Rosen, A. Regev, and S. A. Teichmann, "Single-cell transcriptomics to explore the immune system in health and disease," *Science,* vol. 358, no. 6359, pp. 58, 2017.

[44] S. P. Perfetto, P. K. Chattopadhyay, and M. Roederer, "Seventeen-colour flow cytometry: unravelling the immune system," *Nature Reviews Immunology,* vol. 4, no. 8, pp. 648-655, 2004/08/01, 2004.

[45] D. Chaussabel, V. Pascual, and J. Banchereau, "Assessing the human immune system through blood transcriptomics," *BMC Biology,* vol. 8, no. 1, pp. 84, 2010/07/01, 2010.

[46] L. M. Lepone, R. N. Donahue, I. Grenga, S. Metenou, J. Richards, C. R. Heery, R. A. Madan, J. L. Gulley, and J. Schlom, "Analyses of 123 Peripheral Human Immune Cell Subsets: Defining Differences with Age and between Healthy Donors and Cancer Patients Not Detected in Analysis of Standard Immune Cell Types," *J Circ Biomark,* vol. 5, pp. 5, Jan-Dec, 2016.

[47] S. He, L.-H. Wang, Y. Liu, Y.-Q. Li, H.-T. Chen, J.-H. Xu, W. Peng, G.-W. Lin, P.-P. Wei, and B. Li, "Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs," *Genome biology,* vol. 21, no. 1, pp. 1-34, 2020.

[48] A.-C. Villani, R. Satija, G. Reynolds, S. Sarkizova, K. Shekhar, J. Fletcher, M. Griesbeck, A. Butler, S. Zheng, and S. Lazo, "Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors," *Science,* vol. 356, no. 6335, 2017.

[49] J. Chen, F. Cheung, R. Shi, H. Zhou, and W. Lu, "PBMC fixation and processing for Chromium single-cell RNA sequencing," *Journal of translational medicine,* vol. 16, no. 1, pp. 1-11, 2018.

[50] J. Yang, N. Diaz, J. Adelsberger, X. Zhou, R. Stevens, A. Rupert, J. A. Metcalf, M. Baseler, C. Barbon, and T. Imamichi, "The effects of storage temperature on PBMC gene expression," *BMC immunology,* vol. 17, no. 1, pp. 1-15, 2016.

[51] R. Massoni-Badosa, G. Iacono, C. Moutinho, M. Kulis, N. Palau, D. Marchese, J. Rodríguez-Ubreva, E. Ballestar, G. Rodriguez-Esteban, and S. Marsal, "Sampling time-dependent artifacts in single-cell genomics studies," *Genome biology,* vol. 21, no. 1, pp. 1-16, 2020.

[52] W. C. Olson, M. E. Smolkin, E. M. Farris, R. J. Fink, A. R. Czarkowski, J. H. Fink, K. A. Chianese-Bullock, and C. L. Slingluff, "Shipping blood to a central laboratory in multicenter clinical trials: effect of ambient temperature on specimen temperature, and effects of temperature on mononuclear cell yield, viability and immunologic function," *Journal of translational medicine,* vol. 9, no. 1, pp. 1-13, 2011.

[53] W. C. Olson, M. E. Smolkin, E. M. Farris, R. J. Fink, A. R. Czarkowski, J. H. Fink, K. A. Chianese-Bullock, and C. L. Slingluff, "Shipping blood to a central laboratory in multicenter clinical trials: effect of ambient temperature on specimen temperature, and effects of temperature on mononuclear cell yield, viability and immunologic function," *Journal of translational medicine,* vol. 9, no. 1, pp. 26, 2011.

[54] C. Ducar, D. Smith, C. Pinzon, M. Stirewalt, C. Cooper, M. J. McElrath, and J. Hural, "Benefits of a comprehensive quality program for cryopreserved PBMC covering 28 clinical trials sites utilizing an integrated, analytical web-based portal," *J Immunol Methods,* vol. 409, pp. 9-20, Jul, 2014.

[55] O. Rozenblatt-Rosen, J. W. Shin, J. E. Rood, A. Hupalowska, A. Regev, and H. Heyn, "Building a high-quality human cell atlas," *Nature Biotechnology,* vol. 39, no. 2, pp. 149-153, 2021.

[56] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, and A. R. Green, "SC3: consensus clustering of single-cell RNA-seq data," *Nature methods,* vol. 14, no. 5, pp. 483-486, 2017.

[57] L. Chen, Y. Zhai, Q. He, W. Wang, and M. Deng, “Integrating deep supervised, self-supervised and unsupervised learning for single-cell RNA-seq clustering and annotation,” *Genes,* vol. 11, no. 7, pp. 792, 2020.

[58] W. Hou, Z. Ji, H. Ji, and S. C. Hicks, “A systematic evaluation of single-cell RNA-sequencing imputation methods,” *Genome Biology,* vol. 21, no. 1, pp. 218, 2020/08/27, 2020.

[59] Y. Luning, Y. Zhang, N. Mitic, D. B. Keskin, G. L. ZHANG, L. Chitkushev, and V. Brusic, "Prediction of PBMC Cell Types Using scRNAseq Reference Profiles." pp. 1324-1328.

[60] R. Kramer, J. Mehtonen, G. González, V. Hautamäki, and M. Heinäniemi, “SISUA: Semi-Supervised Generative Autoencoder for Single Cell Data,” 2019.

[61] J. J. Hopfield, “Artificial neural networks,” *IEEE Circuits and Devices Magazine,* vol. 4, no. 5, pp. 3-10, 1988.

[62] T. Jiang, J. L. Gradus, and A. J. Rosellini, “Supervised machine learning: a brief primer,” *Behavior Therapy,* vol. 51, no. 5, pp. 675-687, 2020.

[63] S. B. Maind, and P. Wankar, “Research paper on basic of artificial neural network,” *International Journal on Recent and Innovation Trends in Computing and Communication,* vol. 2, no. 1, pp. 96-100, 2014.

[64] M. M. Mijwel, “Artificial neural networks advantages and disadvantages,” *Retrieved from LinkedIn https//www. linkedin. com/pulse/artificial-neuralnet Work,* 2018.

[65] J. Zhong, R. A. SHAIKH, W. Haoguo, L. Xin, C. Zhiwei, L. T. CHITKUSHEV, G. Zhang, D. B. Keskin, and V. Brusic, "Classification of PBMC cell types using scRNAseq, ANN, and incremental learning." pp. 1351-1355.

[66] I. G. Solman, L. K. Blum, J. A. Burger, T. J. Kipps, J. P. Dean, D. F. James, and A. Mongan, “Impact of long-term ibrutinib treatment on circulating immune cells in previously untreated chronic lymphocytic leukemia,” *Leukemia Research,* vol. 102, pp. 106520, 2021.

[67] M. J. Stubbington, O. Rozenblatt-Rosen, A. Regev, and S. A. Teichmann, “Single-cell transcriptomics to explore the immune system in health and disease,” *Science,* vol. 358, no. 6359, pp. 58-63, 2017.

[68] G. Moncunill, A. Scholzen, M. Mpina, A. Nhabomba, A. B. Hounkpatin, L. Osaba, R. Valls, J. J. Campo, H. Sanz, and C. Jairoce, “Antigen-stimulated PBMC transcriptional protective signatures for malaria immunization,” *Science translational medicine,* vol. 12, no. 543, 2020.

[69] Y. Cai, Y. Dai, Y. Wang, Q. Yang, J. Guo, C. Wei, W. Chen, H. Huang, J. Zhu, and C. Zhang, “Single-cell transcriptomics of blood reveals a natural killer cell subset depletion in tuberculosis,” *EBioMedicine,* vol. 53, pp. 102686, 2020.

[70] A. Sadanandam, T. Bopp, S. Dixit, D. J. Knapp, C. P. Emperumal, P. Vergidis, K. Rajalingam, A. Melcher, and N. Kannan, “A blood transcriptome-based analysis of disease progression, immune regulation, and symptoms in coronavirus-infected patients,” *Cell death discovery,* vol. 6, no. 1, pp. 1-14, 2020.

[71] J. I. Griffiths, P. Wallet, L. T. Pflieger, D. Stenehjem, X. Liu, P. A. Cosgrove, N. A. Leggett, J. A. McQuerry, G. Shrestha, and M. Rossetti, “Circulating immune cell phenotype dynamics reflect the strength of tumor–immune cell interactions in patients during immunotherapy,” *Proceedings of the National Academy of Sciences,* vol. 117, no. 27, pp. 16072-16082, 2020.

[72] V. Chilunda, P. Martinez-Aguado, L. C. Xia, L. Cheney, A. Murphy, V. Veksler, V. Ruiz, T. M. Calderon, and J. W. Berman, "Transcriptional changes in CD16+ monocytes may contribute to the pathogenesis of COVID-19," *Frontiers in Immunology,* vol. 12, pp. 1925, 2021.

[73] E. Vadillo, K. Taniguchi-Ponciano, C. Lopez-Macias, R. Carvente-Garcia, H. Mayani, E. Ferat-Osorio, G. Flores-Padilla, J. Torres, C. R. Gonzalez-Bonilla, and A. Majluf, "A shift towards an immature myeloid profile in peripheral blood of critically Ill COVID-19 patients," *Archives of Medical Research,* vol. 52, no. 3, pp. 311-323, 2021.

[74] Y. Xiong, Y. Liu, L. Cao, D. Wang, M. Guo, A. Jiang, D. Guo, W. Hu, J. Yang, and Z. Tang, "Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients," *Emerging microbes & infections,* vol. 9, no. 1, pp. 761-770, 2020.

[75] T. Gomes, S. A. Teichmann, and C. Talavera-López, "Immunology driven by large-scale single-cell sequencing," *Trends in immunology,* vol. 40, no. 11, pp. 1011-1021, 2019.

[76] A. F. Rendeiro, T. Krausgruber, N. Fortelny, F. Zhao, T. Penz, M. Farlik, L. C. Schuster, A. Nemc, S. Tasnády, and M. Réti, "Chromatin mapping and single-cell immune profiling define the temporal dynamics of ibrutinib response in CLL," *Nature communications,* vol. 11, no. 1, pp. 1-14, 2020.

[77] F. Horns, C. L. Dekker, and S. R. Quake, "Memory B cell activation, broad anti-influenza antibodies, and bystander activation revealed by single-cell transcriptomics," *Cell reports,* vol. 30, no. 3, pp. 905-913. e6, 2020.

[78] A. Noé, T. N. Cargill, C. M. Nielsen, A. J. Russell, and E. Barnes, "The application of single-cell RNA sequencing in vaccinology," *Journal of Immunology Research,* vol. 2020, 2020.

[79] K. E. Yost, H. Y. Chang, and A. T. Satpathy, "Tracking the immune response with single-cell genomics," *Vaccine,* vol. 38, no. 28, pp. 4487-4490, 2020.

[80] C. K. Brierley, and A. J. Mead, "Single-cell sequencing in hematology," *Current opinion in oncology,* vol. 32, no. 2, pp. 139-145, 2020.

[81] X. Ren, W. Wen, X. Fan, W. Hou, B. Su, P. Cai, J. Li, Y. Liu, F. Tang, and F. Zhang, "COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas," *Cell,* vol. 184, no. 7, pp. 1895-1913. e19, 2021.

[82] B. J. Meckiff, C. Ramírez-Suástegui, V. Fajardo, S. J. Chee, A. Kusnadi, H. Simon, S. Eschweiler, A. Grifoni, E. Pelosi, and D. Weiskopf, "Imbalance of regulatory and cytotoxic SARS-CoV-2-reactive CD4+ T cells in COVID-19," *Cell,* vol. 183, no. 5, pp. 1340-1353. e16, 2020.

[83] A. J. Wilk, A. Rustagi, N. Q. Zhao, J. Roque, G. J. Martínez-Colón, J. L. McKechnie, G. T. Ivison, T. Ranganath, R. Vergara, and T. Hollis, "A single-cell atlas of the peripheral immune response in patients with severe COVID-19," *Nature medicine,* vol. 26, no. 7, pp. 1070-1076, 2020.

[84] J.-Y. Zhang, X.-M. Wang, X. Xing, Z. Xu, C. Zhang, J.-W. Song, X. Fan, P. Xia, J.-L. Fu, and S.-Y. Wang, "Single-cell landscape of immunological responses in patients with COVID-19," *Nature immunology,* vol. 21, no. 9, pp. 1107-1118, 2020.

[85] Q. Cao, S. Wu, C. Xiao, S. Chen, X. Chi, X. Cui, H. Tang, W. Su, Y. Zheng, and J. Zhong, "Integrated single-cell analysis revealed immune dynamics during Ad5-nCoV immunization," *Cell Discovery,* vol. 7, no. 1, pp. 1-17, 2021.

[86] Y. Zheng, X. Liu, W. Le, L. Xie, H. Li, W. Wen, S. Wang, S. Ma, Z. Huang, and J. Ye, "A human circulating immune cell landscape in aging and COVID-19," *Protein & cell,* vol. 11, no. 10, pp. 740-770, 2020.

[87] S. W. Kazer, T. P. Aicher, D. M. Muema, S. L. Carroll, J. Ordovas-Montanes, V. N. Miao, A. A. Tu, C. G. Ziegler, S. K. Nyquist, and E. B. Wong, "Integrated single-cell analysis of multicellular immune dynamics during hyperacute HIV-1 infection," *Nature medicine,* vol. 26, no. 4, pp. 511-518, 2020.

[88] A. M. Ranzoni, P. M. Strzelecka, and A. Cvejic, "Application of single-cell RNA sequencing methodologies in understanding haematopoiesis and immunology," *Essays in biochemistry,* vol. 63, no. 2, pp. 217-225, 2019.

[89] J. Acosta, D. Ssozi, and P. van Galen, "Single-Cell RNA Sequencing to Disentangle the Blood System," *Arteriosclerosis, Thrombosis, and Vascular Biology,* vol. 41, no. 3, pp. 1012-1018, 2021.

[90] S. Watcham, I. Kucinski, and B. Gottgens, "New insights into hematopoietic differentiation landscapes from single-cell RNA sequencing," *Blood, The Journal of the American Society of Hematology,* vol. 133, no. 13, pp. 1415-1426, 2019.

[91] L. Hérault, M. Poplineau, A. Mazuel, N. Platet, É. Remy, and E. Duprez, "Single-cell RNA-seq reveals a concomitant delay in differentiation and cell cycle of aged hematopoietic stem cells," *BMC biology,* vol. 19, no. 1, pp. 1-20, 2021.

[92] P. M. Strzelecka, and F. Damm, "Haematopoietic ageing through the lens of single-cell technologies," *Disease models & mechanisms,* vol. 14, no. 1, pp. dmm047340, 2021.

[93] J. Villar, and E. Segura, "Decoding the heterogeneity of human dendritic cell subsets," *Trends in Immunology,* 2020.

[94] D. Zemmour, E. Kiner, and C. Benoist, "CD4+ teff cell heterogeneity: the perspective from single-cell transcriptomics," *Current opinion in immunology,* vol. 63, pp. 61-67, 2020.

[95] H. Chen, F. Ye, and G. Guo, "Revolutionizing immunology with single-cell RNA sequencing," *Cellular & molecular immunology,* vol. 16, no. 3, pp. 242-249, 2019.

[96] D. Arendt, J. M. Musser, C. V. Baker, A. Bergman, C. Cepko, D. H. Erwin, M. Pavlicev, G. Schlosser, S. Widder, and M. D. Laubichler, "The origin and evolution of cell types," *Nature Reviews Genetics,* vol. 17, no. 12, pp. 744-757, 2016.

[97] Y. Panina, P. Karagiannis, A. Kurtz, G. N. Stacey, and W. Fujibuchi, "Human Cell Atlas and cell-type authentication for regenerative medicine," *Experimental & Molecular Medicine,* vol. 52, no. 9, pp. 1443-1451, 2020.

[98] A. Bernard, L. Boumsell, J. Dausset, C. Milstein, and S. F. Schlossman, *Leucocyte Typing: Human Leucocyte Differentiation Antigens Detected by Monoclonal Antibodies. Specification-Classification-Nomenclature/Typage leucocytaire Antigenes de differenciation leucocytaire humains reveles par lesanticorps monoclonaux: Rapports des etudes com*: Springer Science & Business Media, 2013.

[99] M. N. Bernstein, Z. Ma, M. Gleicher, and C. N. Dewey, "CellO: Comprehensive and hierarchical cell type classification of human cells with the Cell Ontology," *Iscience,* vol. 24, no. 1, pp. 101913, 2021.

[100] L. Ziegler-Heitbrock, P. Ancuta, S. Crowe, M. Dalod, V. Grau, D. N. Hart, P. J. Leenen, Y.-J. Liu, G. MacPherson, and G. J. Randolph, "Nomenclature of monocytes and dendritic cells in blood," *Blood, The Journal of the American Society of Hematology,* vol. 116, no. 16, pp. e74-e80, 2010.

[101] L. Zappia, B. Phipson, and A. Oshlack, "Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database," *PLoS computational biology,* vol. 14, no. 6, pp. e1006245, 2018.
[102] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, "Integrating single-cell transcriptomic data across different conditions, technologies, and species," *Nature biotechnology,* vol. 36, no. 5, pp. 411-420, 2018.
[103] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, "Bayesian approach to single-cell differential expression analysis," *Nature methods,* vol. 11, no. 7, pp. 740-742, 2014.
[104] Z. Miao, K. Deng, X. Wang, and X. Zhang, "DEsingle for detecting three types of differential expression in single-cell RNA-seq data," *Bioinformatics,* vol. 34, no. 18, pp. 3223-3224, 2018.
[105] T. Wang, and S. Nabavi, "SigEMD: A powerful method for differential gene expression analysis in single-cell RNA sequencing data," *Methods,* vol. 145, pp. 25-32, 2018.
[106] Z. Wu, Y. Zhang, M. L. Stitzel, and H. Wu, "Two-phase differential expression analysis for single cell RNA-seq," *Bioinformatics,* vol. 34, no. 19, pp. 3340-3348, 2018.
[107] M. Sekula, J. Gaskins, and S. Datta, "Detection of differentially expressed genes in discrete single-cell RNA sequencing data using a hurdle model with correlated random effects," *Biometrics,* vol. 75, no. 4, pp. 1051-1062, 2019.
[108] C. Ye, T. P. Speed, and A. Salim, "DECENT: differential expression with capture efficiency adjustmeNT for single-cell RNA-seq data," *Bioinformatics,* vol. 35, no. 24, pp. 5155-5162, 2019.
[109] D. Aran, A. P. Looney, L. Liu, E. Wu, V. Fong, A. Hsu, S. Chak, R. P. Naikawadi, P. J. Wolters, and A. R. Abate, "Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage," *Nature immunology,* vol. 20, no. 2, pp. 163-172, 2019.
[110] V. Y. Kiselev, A. Yiu, and M. Hemberg, "scmap: projection of single-cell RNA-seq data across data sets," *Nature methods,* vol. 15, no. 5, pp. 359-362, 2018.
[111] A. W. Zhang, C. O'Flanagan, E. A. Chavez, J. L. Lim, N. Ceglia, A. McPherson, M. Wiens, P. Walters, T. Chan, and B. Hewitson, "Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling," *Nature methods,* vol. 16, no. 10, pp. 1007-1015, 2019.
[112] Y. Cao, X. Wang, and G. Peng, "SCSA: a cell type annotation tool for single-cell RNA-seq data," *Frontiers in genetics,* vol. 11, pp. 490, 2020.
[113] R. Hou, E. Denisenko, and A. R. Forrest, "scMatch: a single-cell gene expression profile annotation tool using reference datasets," *Bioinformatics,* vol. 35, no. 22, pp. 4688-4695, 2019.
[114] X. Shao, J. Liao, X. Lu, R. Xue, N. Ai, and X. Fan, "scCATCH: automatic annotation on cell types of clusters from single-cell RNA sequencing data," *IScience,* vol. 23, no. 3, pp. 100882, 2020.
[115] S. Domanskyi, A. Szedlak, N. T. Hawkins, J. Wang, G. Paternostro, and C. Piermarocchi, "Polled Digital Cell Sorter (p-DCS): Automatic identification of hematological cell types from single cell RNA-sequencing clusters," *BMC bioinformatics,* vol. 20, no. 1, pp. 1-16, 2019.

[116] K. Sato, K. Tsuyuzaki, K. Shimizu, and I. Nikaido, "CellFishing. jl: an ultrafast and scalable cell search method for single-cell RNA sequencing," *Genome biology,* vol. 20, no. 1, pp. 1-23, 2019.
[117] A. Senabouth, S. W. Lukowski, J. A. Hernandez, S. B. Andersen, X. Mei, Q. H. Nguyen, and J. E. Powell, "ascend: R package for analysis of single-cell RNA-seq data," *GigaScience,* vol. 8, no. 8, pp. giz087, 2019.
[118] P. Lin, M. Troup, and J. W. Ho, "CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data," *Genome biology,* vol. 18, no. 1, pp. 1-11, 2017.
[119] A. T. Lun, K. Bach, and J. C. Marioni, "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts," *Genome biology,* vol. 17, no. 1, pp. 1-14, 2016.
[120] C. Yau, "pcaReduce: hierarchical clustering of single cell transcriptional profiles," *BMC bioinformatics,* vol. 17, no. 1, pp. 1-11, 2016.
[121] S. Aibar, C. B. González-Blas, T. Moerman, H. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, J. Aerts, and J. van den Oord, "SCENIC: single-cell regulatory network inference and clustering," *Nature methods,* vol. 14, no. 11, pp. 1083-1086, 2017.
[122] M. Guo, H. Wang, S. S. Potter, J. A. Whitsett, and Y. Xu, "SINCERA: a pipeline for single-cell RNA-Seq profiling analysis," *PLoS computational biology,* vol. 11, no. 11, pp. e1004575, 2015.
[123] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev, "Spatial reconstruction of single-cell gene expression data," *Nature biotechnology,* vol. 33, no. 5, pp. 495-502, 2015.
[124] F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY: large-scale single-cell gene expression data analysis," *Genome biology,* vol. 19, no. 1, pp. 1-5, 2018.
[125] B. Wang, D. Ramazzotti, L. De Sano, J. Zhu, E. Pierson, and S. Batzoglou, "SIMLR: A tool for large-scale genomic analyses by multi-kernel learning," *Proteomics,* vol. 18, no. 2, pp. 1700232, 2018.
[126] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells," *Nature biotechnology,* vol. 32, no. 4, pp. 381-386, 2014.
[127] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell, "Reversed graph embedding resolves complex single-cell trajectories," *Nature methods,* vol. 14, no. 10, pp. 979-982, 2017.
[128] K. K. Dey, C. J. Hsiao, and M. Stephens, "Visualizing the structure of RNA-seq expression data using grade of membership models," *PLoS genetics,* vol. 13, no. 3, pp. e1006599, 2017.
[129] D. Grün, M. J. Muraro, J.-C. Boisset, K. Wiebrands, A. Lyubimova, G. Dharmadhikari, M. van den Born, J. Van Es, E. Jansen, and H. Clevers, "De novo prediction of stem cell identity using single-cell transcriptome data," *Cell stem cell,* vol. 19, no. 2, pp. 266-277, 2016.
[130] J. S. Herman, and D. Grün, "FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data," *Nature methods,* vol. 15, no. 5, pp. 379-386, 2018.
[131] D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, and A. Van Oudenaarden, "Single-cell messenger RNA sequencing reveals rare intestinal cell types," *Nature,* vol. 525, no. 7568, pp. 251-255, 2015.

[132] L. Yang, J. Liu, Q. Lu, A. D. Riggs, and X. Wu, "SAIC: an iterative clustering approach for analysis of single cell RNA-seq data," *BMC genomics,* vol. 18, no. 6, pp. 9-17, 2017.
[133] H. Zhang, C. A. Lee, Z. Li, J. R. Garbe, C. R. Eide, R. Petegrosso, R. Kuang, and J. Tolar, "A multitask clustering approach for single-cell RNA-seq analysis in recessive dystrophic epidermolysis bullosa," *PLoS computational biology,* vol. 14, no. 4, pp. e1006053, 2018.
[134] Z. Ji, and H. Ji, "TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis," *Nucleic acids research,* vol. 44, no. 13, pp. e117-e117, 2016.
[135] Y. Yang, R. Huh, H. W. Culpepper, Y. Lin, M. I. Love, and Y. Li, "SAFE-clustering: single-cell aggregated (from ensemble) clustering for single-cell RNA-seq data," *Bioinformatics,* vol. 35, no. 8, pp. 1269-1277, 2019.
[136] Z. Zhang, D. Luo, X. Zhong, J. H. Choi, Y. Ma, S. Wang, E. Mahrt, W. Guo, E. W. Stawiski, and Z. Modrusan, "SCINA: a semi-supervised subtyping algorithm of single cells and bulk samples," *Genes,* vol. 10, no. 7, pp. 531, 2019.
[137] T. S. Johnson, T. Wang, Z. Huang, C. Y. Yu, Y. Wu, Y. Han, Y. Zhang, K. Huang, and J. Zhang, "LAmbDA: label ambiguous domain adaptation dataset integration reduces batch effects and improves subtype detection," *Bioinformatics,* vol. 35, no. 22, pp. 4696-4706, 2019.
[138] C. Xu, R. Lopez, E. Mehlman, J. Regier, M. I. Jordan, and N. Yosef, "Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models," *Molecular systems biology,* vol. 17, no. 1, pp. e9620, 2021.
[139] J. C. Kimmel, and D. R. Kelley, "Semi-supervised adversarial neural networks for single-cell classification," *Genome Research,* pp. gr. 268581.120, 2021.
[140] Q. Song, J. Su, and W. Zhang, "scGCN is a graph convolutional networks algorithm for knowledge transfer in single cell omics," *Nature Communications,* vol. 12, no. 1, pp. 1-11, 2021.
[141] H. Li, E. T. Courtois, D. Sengupta, Y. Tan, K. H. Chen, J. J. L. Goh, S. L. Kong, C. Chua, L. K. Hon, and W. S. Tan, "Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors," *Nature genetics,* vol. 49, no. 5, pp. 708-718, 2017.
[142] H. A. Pliner, J. Shendure, and C. Trapnell, "Supervised classification enables rapid annotation of cell atlases," *Nature methods,* vol. 16, no. 10, pp. 983-986, 2019.
[143] F. Ma, and M. Pellegrini, "ACTINN: automated identification of cell types in single cell RNA sequencing," *Bioinformatics,* vol. 36, no. 2, pp. 533-538, 2020.
[144] J. K. de Kanter, P. Lijnzaad, T. Candelli, T. Margaritis, and F. C. Holstege, "CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing," *Nucleic acids research,* vol. 47, no. 16, pp. e95-e95, 2019.
[145] P. Xie, M. Gao, C. Wang, J. Zhang, P. Noel, C. Yang, D. Von Hoff, H. Han, M. Q. Zhang, and W. Lin, "SuperCT: a supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles," *Nucleic acids research,* vol. 47, no. 8, pp. e48-e48, 2019.
[146] J. Zhong, M. Lyu, H. Jin, Z. Cao, L. T. Chitkushev, G. Zhang, D. B. Keskin, and V. Brusic, "Artificial Neural Networks for classification of single cell gene expression," *bioRxiv,* 2021.

[147] J. Alquicira-Hernandez, A. Sathe, H. P. Ji, Q. Nguyen, and J. E. Powell, "scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data," *Genome biology,* vol. 20, no. 1, pp. 1-17, 2019.
[148] L. Michielsen, M. J. Reinders, and A. Mahfouz, "Hierarchical progressive learning of cell identities in single-cell data," *Nature communications,* vol. 12, no. 1, pp. 1-12, 2021.
[149] Y. Tan, and P. Cahan, "SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species," *Cell systems,* vol. 9, no. 2, pp. 207-213. e2, 2019.
[150] Y. Kaymaz, F. Ganglberger, M. Tang, C. Haslinger, F. Fernandez-Albert, N. Lawless, and T. B. Sackton, "HieRFIT: A hierarchical cell type classification tool for projections from complex single-cell atlas datasets," *Bioinformatics,* 2021.
[151] C. Xu, and Z. Su, "Identification of cell types from single-cell transcriptomes using a novel clustering method," *Bioinformatics,* vol. 31, no. 12, pp. 1974-1980, 2015.
[152] Y. Lin, Y. Cao, H. J. Kim, A. Salim, T. P. Speed, D. M. Lin, P. Yang, and J. Y. H. Yang, "scClassify: sample size estimation and multiscale classification of cells using single and multiple reference," *Molecular systems biology,* vol. 16, no. 6, pp. e9389, 2020.
[153] B. Fa, T. Wei, Y. Zhou, L. Johnston, X. Yuan, Y. Ma, Y. Zhang, and Z. Yu, "GapClust is a light-weight approach distinguishing rare cells from voluminous single cell expression profiles," *Nature Communications,* vol. 12, no. 1, pp. 1-11, 2021.
[154] S. Freytag, L. Tian, I. Lönnstedt, M. Ng, and M. Bahlo, "Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data," *F1000Research,* vol. 7, 2018.
[155] A. Duò, M. D. Robinson, and C. Soneson, "A systematic performance evaluation of clustering methods for single-cell RNA-seq data," *F1000Research,* vol. 7, 2018.
[156] X. Zhang, Y. Lan, J. Xu, F. Quan, E. Zhao, C. Deng, T. Luo, L. Xu, G. Liao, and M. Yan, "CellMarker: a manually curated resource of cell markers in human and mouse," *Nucleic acids research,* vol. 47, no. D1, pp. D721-D728, 2019.
[157] B. J. Schmiedel, D. Singh, A. Madrigal, A. G. Valdovino-Gonzalez, B. M. White, J. Zapardiel-Gonzalo, B. Ha, G. Altay, J. A. Greenbaum, and G. McVicker, "Impact of genetic polymorphisms on human immune cell gene expression," *Cell,* vol. 175, no. 6, pp. 1701-1715. e16, 2018.
[158] V. Chandra, S. Bhattacharyya, B. J. Schmiedel, A. Madrigal, C. Gonzalez-Colin, S. Fotsing, A. Crinklaw, G. Seumois, P. Mohammadi, and M. Kronenberg, "Promoter-interacting expression quantitative trait loci are enriched for functional genetic variants," *Nature Genetics,* vol. 53, no. 1, pp. 110-119, 2021.
[159] P. J. Thul, L. Åkesson, M. Wiking, D. Mahdessian, A. Geladaki, H. A. Blal, T. Alm, A. Asplund, L. Björk, and L. M. Breckels, "A subcellular map of the human proteome," *Science,* vol. 356, no. 6340, 2017.
[160] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, and A. Asplund, "Tissue-based map of the human proteome," *Science,* vol. 347, no. 6220, 2015.
[161] H. Consortium, "The human body at cellular resolution: the NIH Human Biomolecular Atlas Program," *Nature,* vol. 574, no. 7777, pp. 187, 2019.
[162] O. Rozenblatt-Rosen, M. J. Stubbington, A. Regev, and S. A. Teichmann, "The Human Cell Atlas: from vision to reality," *Nature News,* vol. 550, no. 7677, pp. 451, 2017.

[163] L. M. Lepone, R. N. Donahue, I. Grenga, S. Metenou, J. Richards, C. R. Heery, R. A. Madan, J. L. Gulley, and J. Schlom, “Analyses of 123 peripheral human immune cell subsets: defining differences with age and between healthy donors and cancer patients not detected in analysis of standard immune cell types,” *Journal of circulating biomarkers,* vol. 5, no. Godište 2016, pp. 5-5, 2016.

[164] D. Osumi-Sutherland, C. Xu, M. Keays, P. V. Kharchenko, A. Regev, E. Lein, and S. A. Teichmann, “Cell types and ontologies of the Human Cell Atlas,” *arXiv preprint arXiv:2106.14443*, 2021.

[165] R. G. Lindeboom, A. Regev, and S. A. Teichmann, “Towards a Human Cell Atlas: Taking Notes from the Past,” *Trends in Genetics*, 2021.

[166] R. Crevel, "Lymphocyte Proliferation," *Encyclopedic Reference of Immunotoxicology*, H.-W. Vohr, ed., pp. 401-405, Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.

[167] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, "Lymphocytes and the cellular basis of adaptive immunity," *Molecular Biology of the Cell. 4th edition*: Garland Science, 2002.

[168] J. R. Chubb, and T. B. Liverpool, “Bursts and pulses: insights from single cell studies into transcriptional mechanisms,” *Current opinion in genetics & development,* vol. 20, no. 5, pp. 478-484, 2010.

[169] R. Hanamsagar, T. Reizis, M. Chamberlain, R. Marcus, F. O. Nestle, E. de Rinaldis, and V. Savova, “An optimized workflow for single-cell transcriptomics and repertoire profiling of purified lymphocytes from clinical samples,” *Scientific reports,* vol. 10, no. 1, pp. 1-15, 2020.

[170] I. Consuegra, C. Rodríguez-Aierbe, I. Santiuste, A. Bosch, R. Martínez-Marín, M. A. Fortuto, T. Díaz, S. Martí, and M. Á. Muñoz-Fernández, “Isolation methods of peripheral blood mononuclear cells in Spanish Biobanks: an overview,” *Biopreservation and biobanking,* vol. 15, no. 4, pp. 305-309, 2017.

[171] C. P. Corkum, D. P. Ings, C. Burgess, S. Karwowska, W. Kroll, and T. I. Michalak, “Immune cell subsets and their gene expression profiles from human PBMC isolated by Vacutainer Cell Preparation Tube (CPT™) and standard density gradient,” *BMC immunology,* vol. 16, no. 1, pp. 1-18, 2015.

[172] C.-C. Hon, J. W. Shin, P. Carninci, and M. J. Stubbington, “The Human Cell Atlas: technical approaches and challenges,” *Briefings in functional genomics,* vol. 17, no. 4, pp. 283-294, 2018.

[173] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, and M. M. Hoffman, “Opportunities and obstacles for deep learning in biology and medicine,” *Journal of The Royal Society Interface,* vol. 15, no. 141, pp. 20170387, 2018.

[174] N. Novershtern, A. Subramanian, L. N. Lawton, R. H. Mak, W. N. Haining, M. E. McConkey, N. Habib, N. Yosef, C. Y. Chang, and T. Shay, “Densely interconnected transcriptional circuits control cell states in human hematopoiesis,” *Cell,* vol. 144, no. 2, pp. 296-309, 2011.

[175] G. Monaco, B. Lee, W. Xu, S. Mustafah, Y. Y. Hwang, C. Carre, N. Burdin, L. Visan, M. Ceccarelli, and M. Poidinger, “RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types,” *Cell reports,* vol. 26, no. 6, pp. 1627-1640. e7, 2019.

[176] X. Xie, M. Liu, Y. Zhang, B. Wang, C. Zhu, C. Wang, Q. Li, Y. Huo, J. Guo, and C. Xu, "Single-cell transcriptomic landscape of human blood cells," *National Science Review,* vol. 8, no. 3, pp. nwaa180, 2021.
[177] Y. Xu, G.-H. Su, D. Ma, Y. Xiao, Z.-M. Shao, and Y.-Z. Jiang, "Technological advances in cancer immunity: from immunogenomics to single-cell analysis and artificial intelligence," *Signal Transduction and Targeted Therapy,* vol. 6, no. 1, pp. 1-23, 2021.
[178] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos, "The role of ontologies in biological and biomedical research: a functional perspective," *Briefings in bioinformatics,* vol. 16, no. 6, pp. 1069-1080, 2015.
[179] M. A. Haendel, C. G. Chute, and P. N. Robinson, "Classification, ontology, and precision medicine," *New England Journal of Medicine,* vol. 379, no. 15, pp. 1452-1462, 2018.
[180] A. D. Diehl, T. F. Meehan, Y. M. Bradford, M. H. Brush, W. M. Dahdul, D. S. Dougall, Y. He, D. Osumi-Sutherland, A. Ruttenberg, and S. Sarntivijai, "The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability," *Journal of biomedical semantics,* vol. 7, no. 1, pp. 1-10, 2016.
[181] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, and J. T. Eppig, "Gene ontology: tool for the unification of biology," *Nature genetics,* vol. 25, no. 1, pp. 25-29, 2000.
[182] "The Gene Ontology resource: enriching a GOld mine," *Nucleic Acids Research,* vol. 49, no. D1, pp. D325-D334, 2021.
[183] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, and C. J. Mungall, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nature biotechnology,* vol. 25, no. 11, pp. 1251-1255, 2007.
[184] C. Trapnell, "Defining cell types and states with single-cell genomics," *Genome research,* vol. 25, no. 10, pp. 1491-1498, 2015.
[185] T. A. Fleisher, W. T. Shearer, H. W. Schroeder, A. J. Frew, and C. M. Weyand, *Clinical Immunology: Principles and Practice*: Elsevier, 2019.
[186] S. L. Smith, P. R. Kennedy, K. B. Stacey, J. D. Worboys, A. Yarwood, S. Seo, E. H. Solloa, B. Mistretta, S. S. Chatterjee, and P. Gunaratne, "Diversity of peripheral blood human NK cells identified by single-cell RNA sequencing," *Blood advances,* vol. 4, no. 7, pp. 1388-1406, 2020.
[187] C. H. Waddington, "Canalization of development and the inheritance of acquired characters," *Nature,* vol. 150, no. 3811, pp. 563-565, 1942.
[188] I. Sanz, C. Wei, S. A. Jenks, K. S. Cashman, C. Tipton, M. C. Woodruff, J. Hom, and F. Lee, "Challenges and opportunities for consistent classification of human B cell and plasma cell populations," *Frontiers in immunology,* vol. 10, pp. 2458, 2019.
[189] H. W. Schroeder Jr, A. Radbruch, and C. Berek, "B-cell development and differentiation," *Clinical immunology*, pp. 107-118. e1: Elsevier, 2019.
[190] A. Stewart, J. C.-F. Ng, G. Wallis, V. Tsioligka, F. Fraternali, and D. K. Dunn-Walters, "Single-cell transcriptomic analyses define distinct peripheral B cell subsets and discrete development pathways," *Frontiers in immunology,* vol. 12, pp. 743, 2021.
[191] M. Perez-Andres, B. Paiva, W. G. Nieto, A. Caraux, A. Schmitz, J. Almeida, R. Vogt Jr, G. Marti, A. Rawstron, and M. Van Zelm, "Human peripheral blood B-cell compartments:

a crossroad in B-cell traffic," *Cytometry Part B: Clinical Cytometry,* vol. 78, no. S1, pp. S47-S60, 2010.
[192] D. Allman, and S. Pillai, "Peripheral B cell subsets," *Current opinion in immunology,* vol. 20, no. 2, pp. 149-157, 2008.
[193] T. W. LeBien, and T. F. Tedder, "B lymphocytes: how they develop and function," *Blood, The Journal of the American Society of Hematology,* vol. 112, no. 5, pp. 1570-1580, 2008.
[194] M. Seifert, and R. Küppers, "Human memory B cells," *Leukemia,* vol. 30, no. 12, pp. 2283-2292, 2016.
[195] E. C. Rosser, and C. Mauri, "Regulatory B cells: origin, phenotype, and function," *Immunity,* vol. 42, no. 4, pp. 607-612, 2015.
[196] A. M. Masci, C. N. Arighi, A. D. Diehl, A. E. Lieberman, C. Mungall, R. H. Scheuermann, B. Smith, and L. G. Cowell, "An improved ontological representation of dendritic cells as a paradigm for all cell types," *BMC bioinformatics,* vol. 10, pp. 70-70, 2009.
[197] A. D. Diehl, A. D. Augustine, J. A. Blake, L. G. Cowell, E. S. Gold, T. A. Gondré-Lewis, A. M. Masci, T. F. Meehan, P. A. Morel, and A. Nijnik, "Hematopoietic cell types: prototype for a revised cell ontology," *Journal of biomedical informatics,* vol. 44, no. 1, pp. 75-79, 2011.
[198] D. E. Lewis, and S. E. Blutt, "Organization of the immune system," *Clinical immunology,* pp. 19-38. e1: Elsevier, 2019.
[199] M. Collin, and V. Bigley, "Human dendritic cell subsets: an update," *Immunology,* vol. 154, no. 1, pp. 3-20, 2018.
[200] C. Yang, J. R. Siebert, R. Burns, Z. J. Gerbec, B. Bonacci, A. Rymaszewski, M. Rau, M. J. Riese, S. Rao, and K.-S. Carlson, "Heterogeneity of human bone marrow and blood natural killer cells defined by single-cell transcriptome," *Nature communications,* vol. 10, no. 1, pp. 1-16, 2019.
[201] A. Oras, A. Peet, T. Giese, V. Tillmann, and R. Uibo, "A study of 51 subtypes of peripheral blood immune cells in newly diagnosed young type 1 diabetes patients," *Clinical & Experimental Immunology,* vol. 198, no. 1, pp. 57-70, 2019.
[202] L. E. Harrington, "T-cell development," *Clinical Immunology,* pp. 119-125. e1: Elsevier, 2019.
[203] J. J. O'Shea, M. Gadina, and R. M. Siegel, "Cytokines and cytokine receptors," *Clinical immunology,* pp. 127-155. e1: Elsevier, 2019.
[204] T. N. Eagar, and S. D. Miller, "Helper T-cell subsets and control of the inflammatory response," *Clinical immunology,* pp. 235-245. e1: Elsevier, 2019.
[205] S. L. Nutt, and N. D. Huntington, "Cytotoxic T lymphocytes and natural killer cells," *Clinical Immunology,* pp. 247-259. e1: Elsevier, 2019.
[206] A. Takeuchi, and T. Saito, "CD4 CTL, a cytotoxic subset of CD4+ T cells, their differentiation and function," *Frontiers in immunology,* vol. 8, pp. 194, 2017.
[207] M. R. Vieyra-Lobato, J. Vela-Ojeda, L. Montiel-Cervantes, R. López-Santiago, and M. C. Moreno-Lafont, "Description of CD8+ regulatory T lymphocytes and their specific intervention in graft-versus-host and infectious diseases, autoimmunity, and cancer," *Journal of immunology research,* vol. 2018, 2018.
[208] L. E. Sjaastad, D. L. Owen, S. I. Tracy, and M. A. Farrar, "Phenotypic and Functional Diversity in Regulatory T Cells," *Frontiers in Cell and Developmental Biology,* pp. 2665, 2021.

[209] M. Gutierrez-Arcelus, N. Teslovich, A. R. Mola, R. B. Polidoro, A. Nathan, H. Kim, S. Hannes, K. Slowikowski, G. F. Watts, and I. Korsunsky, "Lymphocyte innateness defined by transcriptional states reflects a balance between proliferation and effector functions," *Nature communications,* vol. 10, no. 1, pp. 1-15, 2019.

[210] D. R. McDonald, and O. Levy, "Innate immunity," *Clinical Immunology,* pp. 39-53. e1: Elsevier, 2019.

[211] Y. Ding, L. Zhou, Y. Xia, W. Wang, Y. Wang, L. Li, Z. Qi, L. Zhong, J. Sun, and W. Tang, "Reference values for peripheral blood lymphocyte subsets of healthy children in China," *Journal of Allergy and Clinical Immunology,* vol. 142, no. 3, pp. 970-973. e8, 2018.

[212] R. Valiathan, K. Deeb, M. Diamante, M. Ashman, N. Sachdeva, and D. Asthana, "Reference ranges of lymphocyte subsets in healthy adults and adolescents with special mention of T cell maturation subsets in adults of South Florida," *Immunobiology,* vol. 219, no. 7, pp. 487-496, 2014.

[213] K. Jentsch-Ullrich, M. Koenigsmann, M. Mohren, and A. Franke, "Lymphocyte subsets' reference ranges in an age-and gender-balanced population of 100 healthy adults—a monocentric German study," *Clinical immunology,* vol. 116, no. 2, pp. 192-197, 2005.

[214] A. Al-Mawali, A. D. Pinto, R. Al Busaidi, and I. Al-Zakwani, "Lymphocyte subsets: reference ranges in an age-and gender-balanced population of Omani healthy adults," *Cytometry Part A,* vol. 83, no. 8, pp. 739-744, 2013.

[215] K. Smits, G. Pottier, J. Smet, V. Dirix, F. Vermeulen, I. De Schutter, M. Carollo, C. Locht, C. M. Ausiello, and F. Mascart, "Different T cell memory in preadolescents after whole-cell or acellular pertussis vaccination," *Vaccine,* vol. 32, no. 1, pp. 111-8, Dec 17, 2013.

[216] T. M. Edwards, and J. P. Myers, "Environmental exposures and gene regulation in disease etiology," *Environmental health perspectives,* vol. 115, no. 9, pp. 1264-1270, 2007.

[217] D. Li, Y. Yang, Y. Li, X. Zhu, and Z. Li, "Epigenetic regulation of gene expression in response to environmental exposures: from bench to model," *Science of The Total Environment,* pp. 145998, 2021.

[218] A. S. Findley, A. Monziani, A. L. Richards, K. Rhodes, M. C. Ward, C. A. Kalita, A. Alazizi, A. Pazokitoroudi, S. Sankararaman, and X. Wen, "Functional dynamic genetic effects on gene regulation are specific to particular cell types and environmental conditions," *Elife,* vol. 10, pp. e67077, 2021.

[219] H. Morbach, E. Eichhorn, J. Liese, and H. Girschick, "Reference values for B cell subpopulations from infancy to adulthood," *Clinical & Experimental Immunology,* vol. 162, no. 2, pp. 271-279, 2010.

[220] S. Mahapatra, E. M. Mace, C. G. Minard, L. R. Forbes, A. Vargas-Hernandez, T. K. Duryea, G. Makedonas, P. P. Banerjee, W. T. Shearer, and J. S. Orange, "High-resolution phenotyping identifies NK cell subsets that distinguish healthy children from adults," *PloS one,* vol. 12, no. 8, pp. e0181134, 2017.

[221] R. Afshar, B. Medoff, and A. Luster, "Allergic asthma: a tale of many T cells," *Clinical & Experimental Allergy,* vol. 38, no. 12, pp. 1847-1857, 2008.

[222] T. Boonpiyathad, M. Sokolowska, H. Morita, B. Rückert, J. I. Kast, M. Wawrzyniak, A. Sangasapaviliya, P. Pradubpongsa, R. Fuengthong, P. Thantiworasit, S. Sirivichayakul, W. W. Kwok, K. Ruxrungtham, M. Akdis, and C. A. Akdis, "Der p 1-specific regulatory T-cell response during house dust mite allergen immunotherapy," *Allergy,* vol. 74, no. 5, pp. 976-985, May, 2019.

[223] H. Hocini, H. Bonnabau, C. Lacabaratz, C. Lefebvre, P. Tisserand, E. Foucat, J.-D. Lelièvre, O. Lambotte, A. Saez–Cirion, and P. Versmisse, "HIV controllers have low inflammation associated with a strong HIV-specific immune response in blood," *Journal of virology,* vol. 93, no. 10, pp. e01690-18, 2019.

[224] G. Mijnheer, and F. Van Wijk, "T-Cell compartmentalization and functional adaptation in autoimmune inflammation: lessons from pediatric rheumatic diseases," *Frontiers in immunology,* vol. 10, pp. 940, 2019.

[225] C. Yao, S. A. Bora, T. Parimon, T. Zaman, O. A. Friedman, J. A. Palatinus, N. S. Surapaneni, Y. P. Matusov, G. C. Chiang, and A. G. Kassar, "Cell-type-specific immune dysregulation in severely Ill COVID-19 Patients," *Cell reports,* vol. 34, no. 1, pp. 108590, 2021.

[226] Y. Cai, Y. Dai, Y. Wang, Q. Yang, J. Guo, C. Wei, W. Chen, H. Huang, J. Zhu, C. Zhang, W. Zheng, Z. Wen, H. Liu, M. Zhang, S. Xing, Q. Jin, C. G. Feng, and X. Chen, "Single-cell transcriptomics of blood reveals a natural killer cell subset depletion in tuberculosis," *EBioMedicine,* vol. 53, pp. 102686, 2020/03/01/, 2020.

[227] T. Komura, M. Yano, A. Miyake, H. Takabatake, M. Miyazawa, N. Ogawa, A. Seki, M. Honda, T. Wada, and S. Matsui, "Immune condition of colorectal cancer patients featured by serum chemokines and gene expressions of CD4+ cells in blood," *Canadian Journal of Gastroenterology and Hepatology,* vol. 2018, 2018.

[228] A. Prat, A. Navarro, L. Paré, N. Reguart, P. Galván, T. Pascual, A. Martínez, P. Nuciforo, L. Comerma, L. Alos, N. Pardo, S. Cedrés, C. Fan, J. S. Parker, L. Gaba, I. Victoria, N. Viñolas, A. Vivancos, A. Arance, and E. Felip, "Immune-Related Gene Expression Profiling After PD-1 Blockade in Non-Small Cell Lung Carcinoma, Head and Neck Squamous Cell Carcinoma, and Melanoma," *Cancer Res,* vol. 77, no. 13, pp. 3540-3550, Jul 1, 2017.

[229] T. N. Gide, C. Quek, A. M. Menzies, A. T. Tasker, P. Shang, J. Holst, J. Madore, S. Y. Lim, R. Velickovic, M. Wongchenko, Y. Yan, S. Lo, M. S. Carlino, A. Guminski, R. P. M. Saw, A. Pang, H. M. McGuire, U. Palendira, J. F. Thompson, H. Rizos, I. P. D. Silva, M. Batten, R. A. Scolyer, G. V. Long, and J. S. Wilmott, "Distinct Immune Cell Populations Define Response to Anti-PD-1 Monotherapy and Anti-PD-1/Anti-CTLA-4 Combined Therapy," *Cancer Cell,* vol. 35, no. 2, pp. 238-255.e6, Feb 11, 2019.

[230] J. Lin, Y. Lu, B. Wang, P. Jiao, and J. Ma, "Analysis of immune cell components and immune-related gene expression profiles in peripheral blood of patients with type 1 diabetes mellitus," 2021.

[231] A. Forsberg, T. Abrahamsson, L. Nilsson, J. Ernerudh, K. Duchén, and M. Jenmalm, "Changes in peripheral immune populations during pregnancy and modulation by probiotics and ω-3 fatty acids," *Scientific reports,* vol. 10, no. 1, pp. 1-11, 2020.

[232] A. Gautam, D. Donohue, A. Hoke, S. A. Miller, S. Srinivasan, B. Sowe, L. Detwiler, J. Lynch, M. Levangie, and R. Hammamieh, "Investigating gene expression profiles of whole blood and peripheral blood mononuclear cells using multiple collection and processing methods," *PLoS One,* vol. 14, no. 12, pp. e0225137, 2019.

[233] X. Wang, L. Yu, and A. R. Wu, "The effect of methanol fixation on single-cell RNA sequencing data," *BMC genomics,* vol. 22, no. 1, pp. 1-16, 2021.

[234] L.-H. Huang, P.-H. Lin, K.-W. Tsai, L.-J. Wang, Y.-H. Huang, H.-C. Kuo, and S.-C. Li, "The effects of storage temperature and duration of blood samples on DNA and RNA qualities," *PloS one,* vol. 12, no. 9, pp. e0184692, 2017.
[235] M. Djaldetti, and H. Bessler, "High temperature affects the phagocytic activity of human peripheral blood mononuclear cells," *Scandinavian journal of clinical and laboratory investigation,* vol. 75, no. 6, pp. 482-486, 2015.
[236] F. Malentacchi, S. Pizzamiglio, R. Wyrich, P. Verderio, C. Ciniselli, M. Pazzagli, and S. Gelmini, "Effects of transport and storage conditions on gene expression in blood samples," *Biopreservation and biobanking,* vol. 14, no. 2, pp. 122-128, 2016.
[237] A. Germann, Y.-J. Oh, T. Schmidt, U. Schön, H. Zimmermann, and H. von Briesen, "Temperature fluctuations during deep temperature cryopreservation reduce PBMC recovery, viability and T-cell function," *Cryobiology,* vol. 67, no. 2, pp. 193-200, 2013.
[238] C. M. Hope, D. Huynh, Y. Y. Wong, H. Oakey, G. B. Perkins, T. Nguyen, S. Binkowski, M. Bui, A. Y. Choo, and E. Gibson, "Optimization of Blood Handling and Peripheral Blood Mononuclear Cell Cryopreservation of Low Cell Number Samples," *International journal of molecular sciences,* vol. 22, no. 17, pp. 9129, 2021.
[239] H. Yamagata, A. Kobayashi, R. Tsunedomi, T. Seki, M. Kobayashi, K. Hagiwara, C. Chen, S. Uchida, G. Okada, and M. Fuchikami, "Optimized protocol for the extraction of RNA and DNA from frozen whole blood sample stored in a single EDTA tube," *Scientific reports,* vol. 11, no. 1, pp. 1-10, 2021.
[240] A. Box, M. DeLay, S. Tighe, S. V. Chittur, A. Bergeron, M. Cochran, P. Lopez, E. M. Meyer, A. Saluk, and S. Thornton, "Evaluating the effects of cell sorting on gene Expression," *Journal of biomolecular techniques: JBT,* vol. 31, no. 3, pp. 100, 2020.
[241] G. M. Richardson, J. Lannigan, and I. G. Macara, "Does FACS perturb gene expression?," *Cytometry Part A,* vol. 87, no. 2, pp. 166-175, 2015.
[242] G. Pfister, S. M. Toor, V. S. Nair, and E. Elkord, "An evaluation of sorter induced cell stress (SICS) on peripheral blood mononuclear cells (PBMCs) after different sort conditions-Are your sorted cells getting SICS?," *Journal of Immunological Methods,* vol. 487, pp. 112902, 2020.
[243] N. Beliakova-Bethell, M. Massanella, C. White, S. M. Lada, P. Du, F. Vaida, J. Blanco, C. A. Spina, and C. H. Woelk, "The effect of cell subset isolation method on gene expression in leukocytes," *Cytometry Part A,* vol. 85, no. 1, pp. 94-104, 2014.
[244] C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, and W. Enard, "Comparative analysis of single-cell RNA sequencing methods," *Molecular cell,* vol. 65, no. 4, pp. 631-643. e4, 2017.
[245] A. Senabouth, S. Andersen, Q. Shi, L. Shi, F. Jiang, W. Zhang, K. Wing, M. Daniszewski, S. W. Lukowski, and S. S. Hung, "Comparative performance of the BGI and Illumina sequencing technology for single-cell RNA-sequencing," *NAR genomics and bioinformatics,* vol. 2, no. 2, pp. lqaa034, 2020.
[246] N. Stoler, and A. Nekrutenko, "Sequencing error profiles of Illumina sequencing instruments," *NAR genomics and bioinformatics,* vol. 3, no. 1, pp. lqab019, 2021.
[247] S. Rizzetto, A. A. Eltahla, P. Lin, R. Bull, A. R. Lloyd, J. W. Ho, V. Venturi, and F. Luciani, "Impact of sequencing depth and read length on single cell RNA sequencing data of T cells," *Scientific reports,* vol. 7, no. 1, pp. 1-11, 2017.

[248] Y. Zhang, Y. Luning, and V. Brusic, "Automation of Gene Expression Profile Analysis in Single Cell Data." pp. 1329-1334.
[249] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research,* vol. 12, pp. 2825-2830, 2011.
[250] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980,* 2014.
[251] E. Clough, and T. Barrett, "The gene expression omnibus database In: Statistical Genomics," New York, NY: Humana Press, 2016.
[252] M. Guilliams, F. Ginhoux, C. Jakubzick, S. H. Naik, N. Onai, B. U. Schraml, E. Segura, R. Tussiwand, and S. Yona, "Dendritic cells, monocytes and macrophages: a unified nomenclature based on ontogeny," *Nature Reviews Immunology,* vol. 14, no. 8, pp. 571-578, 2014.
[253] R. Ade, and P. Deshmukh, "Methods for incremental learning: a survey," *International Journal of Data Mining & Knowledge Management Process,* vol. 3, no. 4, pp. 119, 2013.
[254] A. Barbarin, E. Cayssials, F. Jacomet, N. G. Nunez, S. Basbous, L. Lefèvre, M. Abdallah, N. Piccirilli, B. Morin, V. Lavoue, V. Catros, E. Piaggio, A. Herbelin, and J.-M. Gombert, "Phenotype of NK-Like CD8(+) T Cells with Innate Features in Humans and Their Relevance in Cancer Diseases," *Frontiers in Immunology,* vol. 8, no. 316, 2017-March-27, 2017.
[255] N. Borcherding, A. P. Voigt, V. Liu, B. K. Link, W. Zhang, and A. Jabbari, "Single-cell profiling of cutaneous T-cell lymphoma reveals underlying heterogeneity associated with disease progression," *Clinical Cancer Research,* vol. 25, no. 10, pp. 2996-3005, 2019.
[256] L. Brockmann, S. Soukou, B. Steglich, P. Czarnewski, L. Zhao, S. Wende, T. Bedke, C. Ergen, C. Manthey, and T. Agalioti, "Molecular and functional heterogeneity of IL-10-producing CD4+ T cells," *Nature communications,* vol. 9, no. 1, pp. 1-14, 2018.
[257] N. Ranu, A.-C. Villani, N. Hacohen, and P. C. Blainey, "Targeting individual cells by barcode in pooled sequence libraries," *Nucleic acids research,* vol. 47, no. 1, pp. e4-e4, 2019.
[258] C. Goudot, A. Coillard, A.-C. Villani, P. Gueguen, A. Cros, S. Sarkizova, T.-L. Tang-Huau, M. Bohec, S. Baulande, and N. Hacohen, "Aryl hydrocarbon receptor controls monocyte differentiation into dendritic cells versus macrophages," *Immunity,* vol. 47, no. 3, pp. 582-596. e6, 2017.
[259] A. S. W. Davis, H. N. Roozen, M. J. Dufort, H. A. DeBerg, M. A. Delaney, F. Mair, J. R. Erickson, C. K. Slichter, J. D. Berkson, and A. M. Klock, "The human tissue-resident CCR5+ T cell compartment maintains protective and functional properties during inflammation," *Science translational medicine,* vol. 11, no. 521, 2019.
[260] C. R. Kleiveland, "Peripheral Blood Mononuclear Cells," *The Impact of Food Bioactives on Health: in vitro and ex vivo models*, K. Verhoeckx, P. Cotter, I. López-Expósito, C. Kleiveland, T. Lea, A. Mackie, T. Requena, D. Swiatecka and H. Wichers, eds., pp. 161-167, Cham: Springer International Publishing, 2015.
[261] S. C. Technologies., *Frequencies of cell types in human peripheral blood.* , STEMCELL Technologies Inc., www.stemcell.com, 2019.
[262] D. Krijgsman, N. L. de Vries, A. Skovbo, M. N. Andersen, M. Swets, E. Bastiaannet, A. L. Vahrmeijer, C. J. van de Velde, M. H. Heemskerk, and M. Hokland, "Characterization

of circulating T-, NK-, and NKT cell subsets in patients with colorectal cancer: the peripheral blood immune cell profile," *Cancer Immunology, Immunotherapy,* vol. 68, no. 6, pp. 1011-1024, 2019.

[263] E. F. McKinney, I. Cuthbertson, K. M. Harris, D. E. Smilek, C. Connor, G. Manferrari, E. J. Carr, S. S. Zamvil, and K. G. Smith, "A CD8+ NK cell transcriptomic signature associated with clinical outcome in relapsing remitting multiple sclerosis," *Nature communications,* vol. 12, no. 1, pp. 1-9, 2021.

[264] K. Xie, Y. Huang, F. Zeng, Z. Liu, and T. Chen, "scAIDE: clustering of large-scale single-cell RNA-seq data reveals putative and rare cell types," *NAR genomics and bioinformatics,* vol. 2, no. 4, pp. lqaa082, 2020.

[265] X. Chen, F.-X. Wu, J. Chen, and M. Li, "DoRC: Discovery of rare cells from ultra-large scRNA-seq data." pp. 111-116.

[266] M. Gry, R. Rimini, S. Strömberg, A. Asplund, F. Pontén, M. Uhlén, and P. Nilsson, "Correlations between RNA and protein expression profiles in 23 human cell lines," *BMC genomics,* vol. 10, no. 1, pp. 1-14, 2009.

[267] L. R. Olsen, B. Campos, O. Winther, D. C. Sgroi, B. L. Karger, and V. Brusic, "Tumor antigens as proteogenomic biomarkers in invasive ductal carcinomas," *BMC medical genomics,* vol. 7, no. 3, pp. 1-13, 2014.

[268] H. Yu, D. C. Samuels, Y.-y. Zhao, and Y. Guo, "Architectures and accuracy of artificial neural network for disease classification from omics data," *BMC genomics,* vol. 20, no. 1, pp. 1-12, 2019.

[269] T. Borovicka, M. Jirina Jr, P. Kordik, and M. Jirina, "Selecting representative data sets," *Advances in data mining knowledge discovery and applications,* vol. 12, pp. 43-70, 2012.

[270] T.-L. Tang-Huau, P. Gueguen, C. Goudot, M. Durand, M. Bohec, S. Baulande, B. Pasquier, S. Amigorena, and E. Segura, "Human in vivo-generated monocyte-derived dendritic cells and macrophages cross-present antigens through a vacuolar pathway," *Nature Communications,* vol. 9, no. 1, pp. 2570, 2018/07/02, 2018.

[271] A. Baratloo, M. Hosseini, A. Negida, and G. El Ashal, "Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity," *Emerg (Tehran),* vol. 3, no. 2, pp. 48-9, Spring, 2015.

[272] R. Trevethan, "Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice," *Frontiers in Public Health,* vol. 5, 2017-November-20, 2017.

[273] D. K. Cole, B. Laugel, M. Clement, D. A. Price, L. Wooldridge, and A. K. Sewell, "The molecular determinants of CD 8 co-receptor function," *Immunology,* vol. 137, no. 2, pp. 139-148, 2012.

APPENDICES

Appendix 1 Publications and Presentations Arising from This Thesis

○ PAPERS

- J. Zhong, R. A. Shaikh, H. Wu, X. Lin, Z. Cao, L. T. Chitkushev, G. Zhang, D. B. Keskin, and V. Brusic, "Classification of PBMC cell types using scRNAseq, ANN, and incremental learning," IEEE Int. Conf. Bioinform. Biomed., pp. 1351-1355, 2020.

- R. A. Shaikh, J. Zhong, M. Lyu, S. Lin, D. B. Keskin, G. Zhang, L. Chitkushev, and V. Brusic, "Classification of Five Cell Types from PBMC Samples using Single Cell Transcriptomics and Artificial Neural Networks," IEEE Int. Conf. Bioinform. Biomed., pp. 2207-2213, 2019.

○ PREPRINT MANUSCRIPT

- J. Zhong, M. Lyu, H. Jin, Z. Cao, L. T. Chitkushev, G. Zhang, D. B. Keskin, and V. Brusic, "Artificial Neural Networks for classification of single cell gene expression," bioRxiv, 2021.

○ ORAL PRESENTATINS

- Classification of PBMC cell types using scRNA-seq, ANN, and incremental learning, International Conference on Bioinformatics and Biomedicine (BIBM) 2020, 16th-19th December 2020, Seoul, South Korea - online virtually.

- Classification of Cells Using Single Cell Transcriptomics Data and Artificial Neural Networks, The 6th UNNC Postgraduate Research Conference, 20th November 2020, Ningbo, China.

○ POSTER PRESENTATIONS

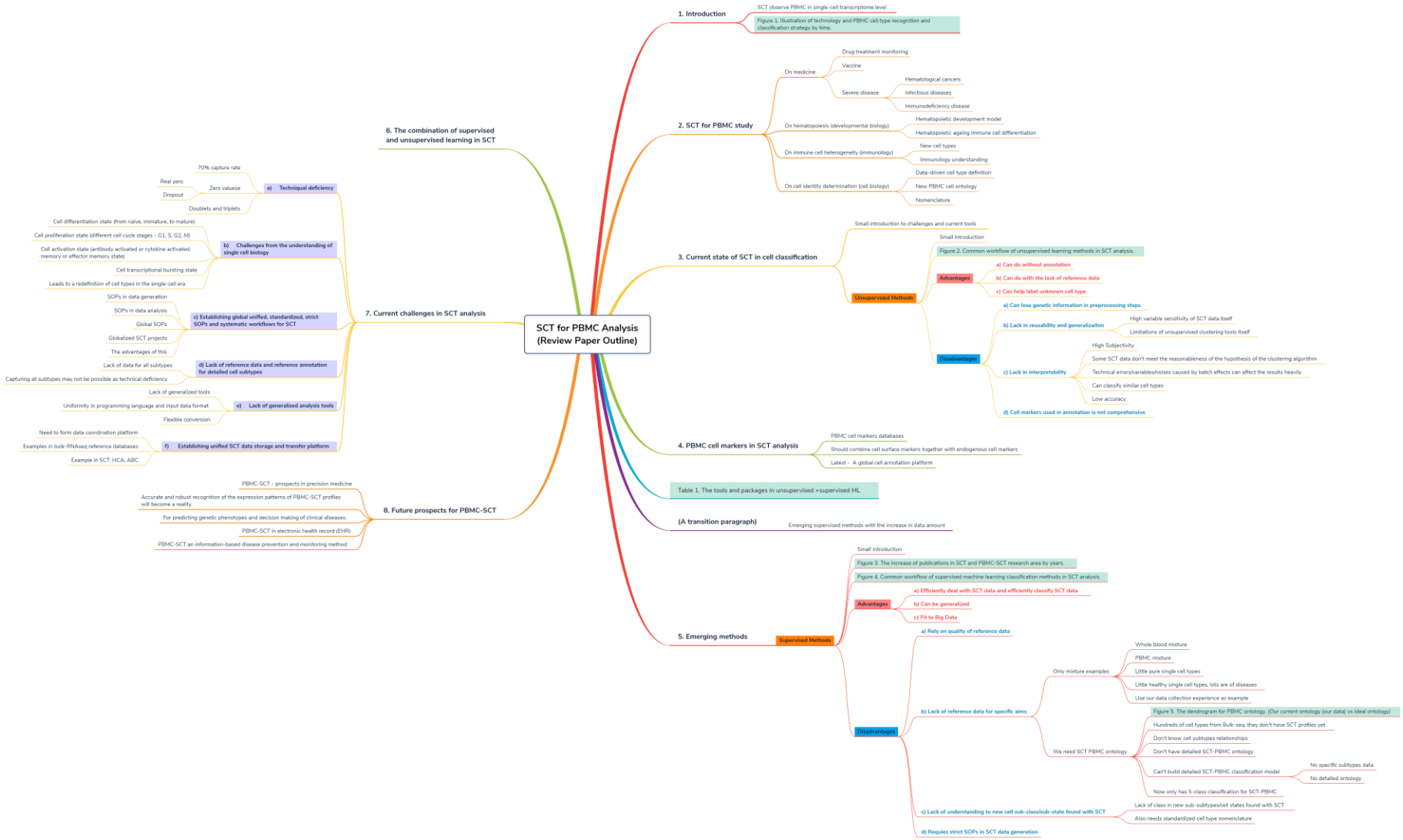
- Artificial Neural Networks for Classification of Single Cell Gene Expression. The 3rd Annual Faculty of Science and Engineering Postgraduate Research Showcase Poster Exhibition, 21st May 2021, Ningbo, China.

- Peripheral Blood Mononuclear Cell Classification using Single-cell RNA-seq Data and Artificial Neural Networks. The 4th Annual Faculty of Science and Engineering Postgraduate Research Showcase Poster Exhibition, 10th June 2022, Ningbo, China.

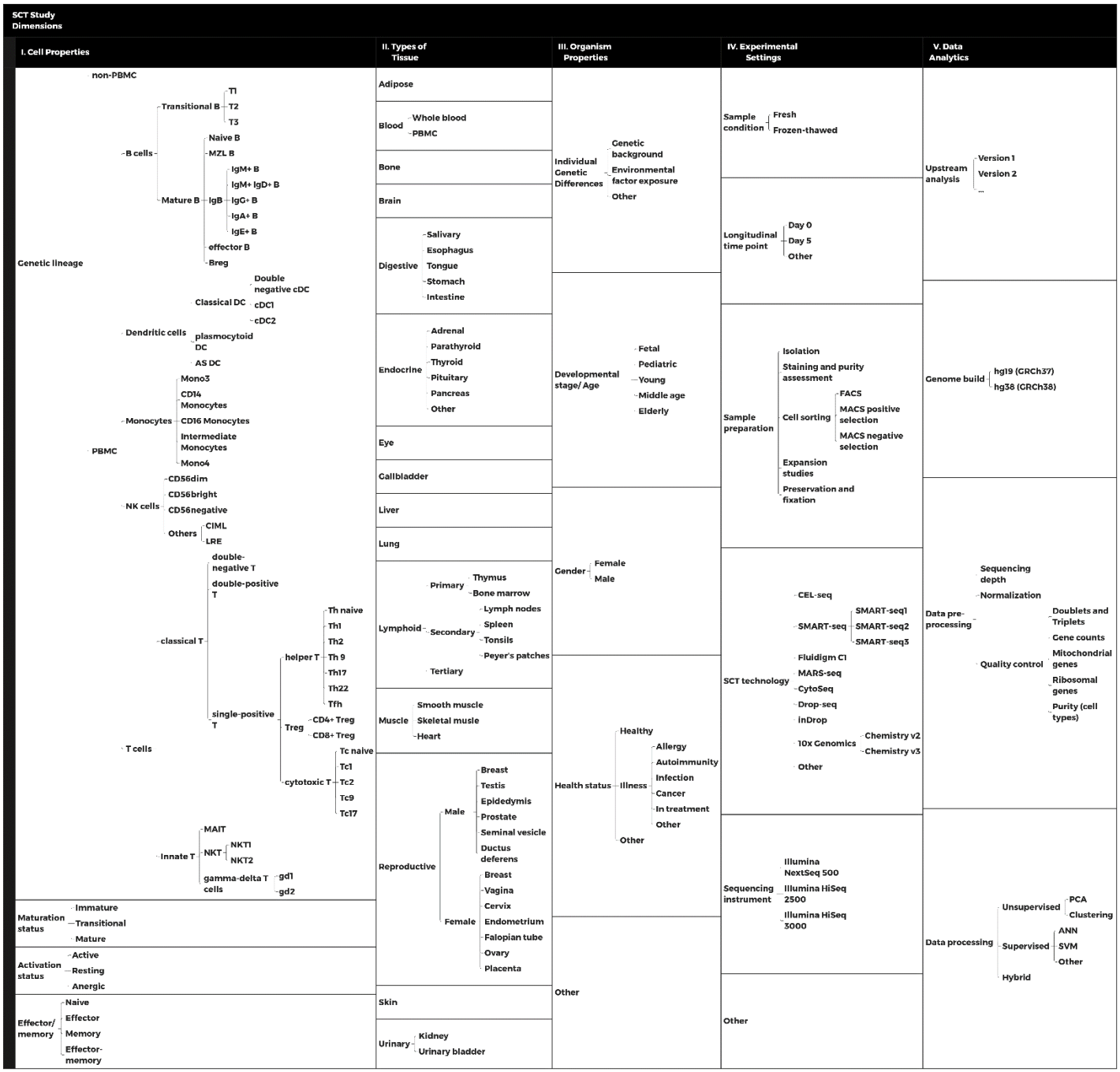
Appendix 2 Reference SCT Datasets

All datasets from this study are available at http://projects.methilab.org/SCTdata/PBMC001

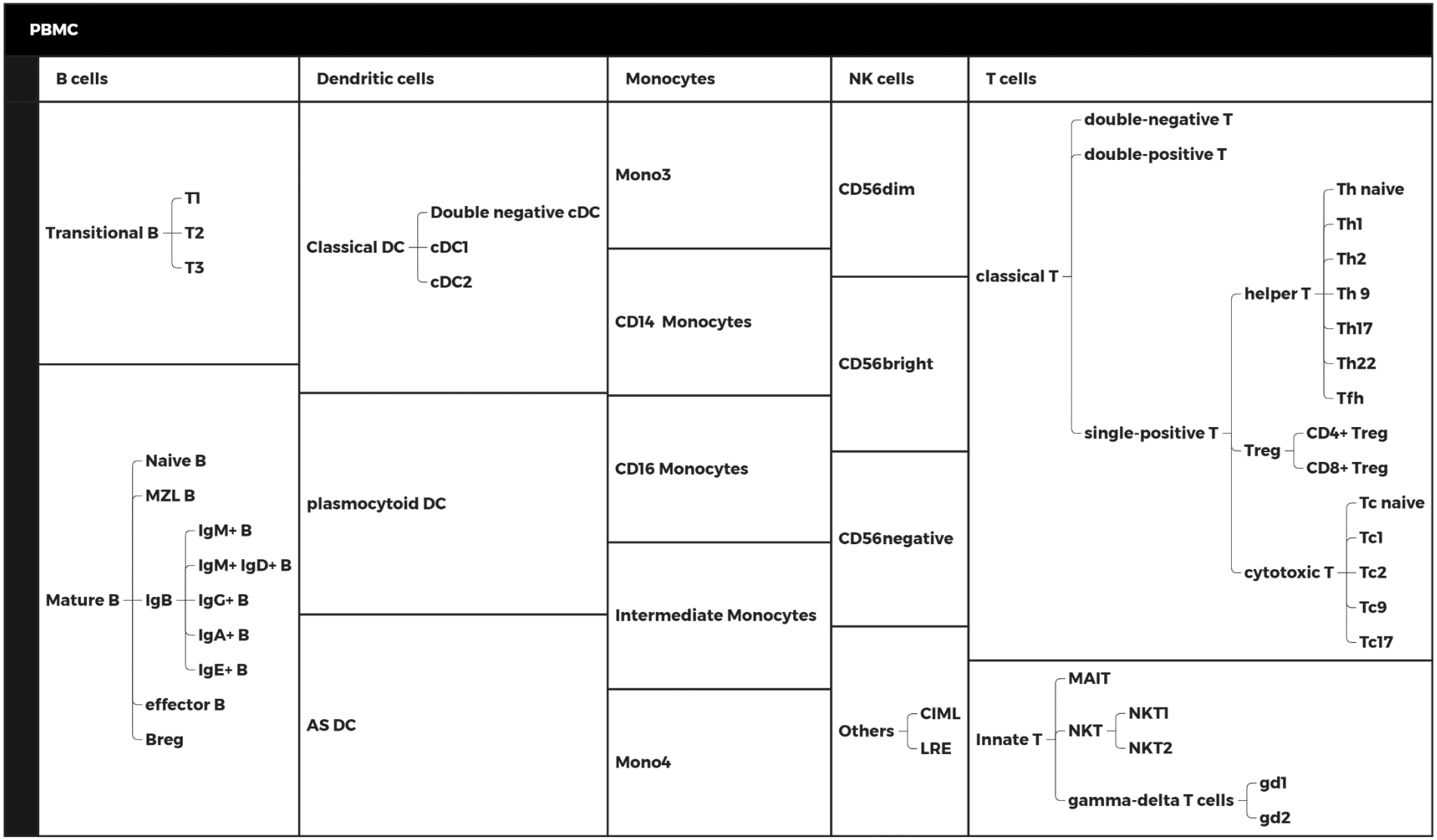
Appendix 3 Outline Graph of the Literature Review



Appendix 4 SCT Study Dimensions



Appendix 5 PBMC Dimensions



Appendix 6 Cell Ontology Construction Metadata (PBMC Section)

PBMC ontology (focused on our data, 3 layers in total - one current layer, one upper, one lower)
v4:
* gray color means not of our data set label directly
*black color shows the cell types represented in our data (described in metadata)
*Frequency in this table refers to it of PBMCs (% of PBMC)
* B, T subset is ordered by maturation status
*remember this table is for PBMC, blood, circulating cells
*Green background color means that of our data sest label directly
*naming tries to be the same with the latest Nomenclature (Nomenclature Committee of the IUIS)
*this version of ontology tries to focus on the checkpoint cell types (parental/children types), precursors or transitional types are not focused

Pa	Circu									Add	Frequency in PBMC (	Phenotype	Core markers	Additional markers	Function/Properties	Not	The labels that we h	Yes	GEO	10X	Bloo	Br	Background information for 5 main types in PB	Reference	
											≥ 2 % [1], 0.5-1% [41]		CD3−CD19−CD20−CD14−CD56−HLA−DR+				DC	1		V	V		DCs in blood are less mature and have no dem	[1][2][42][43][67][27]	
													CD3−CD19−CD20−CD14−CD56−HLA−DR+CD123+	CD1c+CD303-										[1][27]	
			Myeloid CD1c+ blood								0.6% [6]		CD1c, CD11c, CD123, CD13, CD33, CD32, CD64, FcεRI, CD2, CD45RO, CD141. A small p	Produce IL-12 upon TLmo										[1][2][3][6][46]	
			Myeloid CD141+ bloo								0.05% [6]		ClecSa, XCR1, CD11c, CD13, CD33	IL-12 and IFN-λ secretema										[1][2][3][6][46]	
		Plasmacytoid CD303+ blood DCs											CD303+ (BDCA-2), CD3−CD19−CD20−CD14−CD56−HLA−DR+CD123+	CD1c-	Upon pathogen encou		plasmacytoid dendr	1			V			[1][2][3][6][46][27]	
Mon											5-10% [1], 10-30% [39], 4.13% [4		CD3−CD19−CD20−CD14+											[27]	
	Classical CD14++	CD16- monocytes									Mean: 17.7% Min: 8.34% Max: 4		CD3−CD19−CD20−CD14++CD16−				CD14+Monocyte	1	GSM(	V	V			[46][27][90]	
	Nonclassical CD										Mean: 1.55% Min: 0.38% Max: 4		CD3−CD19−CD20−CD14+CD16+			proc	CD16+ Monocyte	1			V	V	M16 - which comprise up to 10% of the monoc	[46][27][90]	
	Intermediate CD																							[46]	
Lyma											70 – 90% [39]													[39]	
	B cell (5-10% of P										5-10% of PBMC [1],	CD3−CD19+					CD19+ B cell, B cell	1	GSM(	V	V		Total B cell, Anchor marker CD45, CD subset m	[11][12][13][14][27][41]	
		Immature B																						[41]	
		Transitional B										CD3-/CD19-	IgD+CD27−CD38++CD24++									pre-naive	[11][38][41][66][27]	
			T1 B (T1/T27)										IgD+ CD27- CD38++ CD24++	CD10++/+ IgM++ MTG+ C	Developmental precursor									[41][47]	
			T2 B (T2-MZP (MZ precu										IgD+ CD27- CD38++ CD24++ CD21++	CD10+ IgM++	MZ precursor									[41][47]	
			T3 B										IgD+ CD27-CD38++CD24++CD21+	CD10+/- IgM++MTG+	Developmental precursor or activated naive									[41][47]	
		Naive (mature) B									Mean 5.2% Min 1.53	CD3-/CD19-	CD19+, IgD+, CD27−, CD38+/−, CD24+/−		Naive B cell	1			V			express the BCR as IgM and IgD molecules, rec	[9][10][11][41][47][66][27][90]	
				Resting naive B									IgD+ CD27-CD38++CD24++CD21+	IgM+ MTG-	Antigen inexperienced mature cells									[47]	
				Activated naive B									IgD+ CD27- CD38- CD24- CD21-	IgM+ MTG+ CD95+ CD23	Precursor of short-lived PB and GC reactions									[9][10][41][47]	
				Anergic naive B							<2% of blood B cells,	30% of all N	IgD+ CD27- CD38+/lo CD24+ CD21-	IgMlow/-	Hypo-responsive. Maintenance of tolerance									[47]	
		Circulating marginal zone (MZ) B																						[47]	
		Memory B										CD3-/CD19+	CD 19+, CD 27+, CD38+/−, CD24+/−				Memory B cell	1			V			[9][10][11][41][47][66]	
			Unswitched (MZ-like)										IgDlo CD27+ CD38+/lo CD24+ CD21+	CD1c+ IgM++	Natural memory.MZ equivalent									[41][47]	
					IgM-only and IgM+IgD+ memory B cell																	IgM-only B cells (IgM+IgDlow/-) and IgM+IgD+	[52][47]	
			Pre-switched									CD3-CD19+	IgD- CD27+ CD38+/lo CD24+ CD21+	IgM+	Pre-switch memory; early IgM memory. IgG memory precursor									[47][27]	
			Switched									CD3-CD19+	IgD- CD27+									Switched memory B cells (SM B cell)	[27][41][47]	
					IgG+ memory B cell																	Approximately 10-20% of peripheral blood (PB	[52][47]	
					IgA+ memory B cell																	IgA+ memory B cells account for approximately	[52][47]	
					IgE+ memory B cell																	IgE-expressing memory B cells are hardly dete	[52][47]	
			Switched resting										IgD- CD27+ CD38+/lo CD24+ CD21+	IgG/IgA+ CD95-	Pre-existing memory reservoir									[47]	
			Switched activated										IgD- CD27+ CD38- CD24-CD21-	IgG/IgA+ CD95+ CD86+	Effector memory PB/PC precursor									[47]	
			Atypical tissue-based										IgD- CD27- CD38lo CD24lo CD21-	FcRL4+ IgM/IgG/IgA+ FcR	Mucosal surveillance; exhausted memory; BCR hypo-responsive memory									[47]	
		Double negative (IgD− CD27− DN)									<10% of all PBL CD19+ cells [47]		CD3-CD19+/CD20+/IgD-/CD27-, CD38+/−, CD24+/−											[47][66]	
			DN1										IgD- CD27- CD38- CD24+ CD21+	FcRL4 IgM/IgG/IgA+ FcR	Memory precursors									[47]	
			DN2										IgD- CD27- CD38- CD24- CD21-	IgM/IgG/IgA+ T-bet+ CD2	Extrafollicular ASC precursors									[47]	
													IgD- CD27- CD38- CD24- CD21-	FcRL4+	Atypical/Tissue-based memory									[47]	
		Antibody secreting cells (ASC)											CD19+/-, IgD−, CD27++, CD38++, CD24−											[47]	
			Early plasmablast (PB)										IgD- CD27lo/+ CD38++ CD24-	CD20+/- HLA-DR+ CD138	Naïve and memory-derived PB precursors									[47]	
			PB									CD3-/CD19+	IgD- CD27++ CD38+++ CD24-	CD20- HLA-DR+ CD138- K	Antibody secretion									[47][66][27]	
			Plasma cell (PC, effector									CD19+CD24-	IgD- CD27++ CD38+++ CD24-	CD19+/- CD20- CD138+	Antibody secretion								When B cells develop into plasma cells, they dy	[9][10][11][41][52][54][56][47]	
		Regulatory B cells (Breg)									may account for <20%	of B cells	Multiple surface phenotypes corresponding to		Down regulation of T cell and monocyte inflammatory and autoimmune responses									[47]	
		Regulatory plasma cells (PCreg)											IgD- CD27++ CD38++	CD19+/- CD138+ IL-10+ S	Suppress immune responses including anti-tumor responses									[47]	
		B1 cells											IgD++ CD27+	IgM+ CD43+ CD70- CD11	Production of natural autoantibodies									[47]	
	CD3+ T cell (45-70% [39])										45-70% [39], memo		CD3+, Anchor marker CD45, CD3+/CD19-, T cells (CD45+SSClowCD3+)											[39]	
		αβ T																						[41]	
			TCRαβ+ double-negative																					[11][41]	
			TCRαβ+ double-positive																					[11][41]	
			TCRαβ+ single-positive T																					[41]	
				CD4+ T cell (helper T, 25-60% [1])							25-60% [1][39], 27-5	CD3+CD4+CD8-					CD4+ T cell, CD4+ Th cells, GSM(	V			V	CD4 T cell can be further classified into various	[11][39][69][88]	
					Naive CD4						Mean 25.9% in PB [6	CD3+/CD8-	CD45RA+ CCR7+	CD197+			T4 naive, CD4+ CD45	1		V	V			[11][20][62][66][88][89][27]	Th1
					Activated							CD3+/CD8-	CD4+/CD38+/HLADR+											[66]	Th2
					Effector T							CD3+/CD8-	CD4+/CCR7-/CD45RA+											[66][85]	Th37
					Memory C						Mean 54.6% in PB [62]		CD4+CD45RO+				CD4+ CD45RO+ Mem	1		V				[62][84][88][89]	Th9
											Mean 40.2% in PB [6	CD3+/CD8-	CD45RA- CCR7+/CD4+CD62L+ CD45RA-	CD197+									antigen experienced central memory	[11][20][39][66][88][89][27]	Th17
						Contr																			Th1/17
							EM Th1																		Th22
							EM Th3																		Tfh
						Effect					30.5 ± 13.3 % [89]	CD3+/CD8-	CD45RA- CCR7-	CD197−			T4eff,mem	1			V		effector subtypes that exist in resting or activa	[11][20][39][66][88][89][27]	
							EM Th1																		
							EM Th2																		
							EM Th3																		
						Term					Mean 16.9% in PB [6	CCR7-	CD45RA+ CCR7-/CD4+ CD45RO+ CD45RA+	CD197−										[11][20][88][89][27]	
					Regulatory						5-10% of CD4 T cell	CD4+CD25+	CD14−/CD56−CD3+/CD4+CD8−/CD25+	CD127−			CD4+ CD25+ Treg	1		V				[20][27]	
						aTreg						CD45RO+	CD25++CD45RA (FOXP3lo)				aTreg	1			V			[20][25]	
						rTreg						CD45RO−	CD25++CD45RA+ (FOXP3lo)				rTreg	1			V			[20][25]	
						cTreg						CD45RO+	CD45RO+Foxp3low											[20][25]	
																	Non-classical T	1			V				
																	CD4+, CCR5+ CD69-	1	GSM3209408 [34]						
				CD8+ T cell (Cytotoxic T, 5-30% [1]							5-30% [1][39], 13-23	CD3+CD4-CD8+					CD8+ T cell, Cytotox	1	GSM: V		V		CD8 T cell - to be extremely heterogeneous in m	[39][68]	
					Naive CD8						37.6 ± 19.4 % [89]	CD3+/CD8+	CD14−/CD56−CD3+/CD4−CD8+/CD45RA+CCR7+	CD197+			T8naive, CD8+ CD45	1		V	V			[11][20][66][88][89][27]	Tc1
					Activated							CD3+/CD8+	CD4-/CD38+/HLADR+											[66]	Tc2
					Effector C							CD3+/CD8+	CD4-/CCR7-/CD45RA+											[66][85]	Tc9
					Memory C																			[84][88][89]	Tc17
						Contr					6.25 ± 4.40 % [89]	CD3+/CD8+	CD14−/CD56−CD3+/CD4−CD8+/CD45RA−CCR7+	CD197+										[11][20][66][88][89][27]	Tc22
						CD8 T																			
						Effect					29.9 ± 12.3 % [89]	CD3+/CD8+	CD14−/CD56−CD3+/CD4−CD8+/CD45RA−CCR7−	CD197−			T8eff,mem	1			V			[11][20][66][88][89][27]	
						Term					26.2 ± 18.1 % [89]	CCR7-	CD14−/CD56−CD3+/CD4−CD8+/CD45RA+CCR7−	CD197−										[11][20][88][89][27]	
					RegT (CD8CD8+CD28 - Treg																				
						CD8+ CD28-limited Treg																			
		γδ T																							
			Innate-like T cell																						
				Natural killer T cell								CD56+ CD3+												[20]	
					Invariant	NKT (INKT)											INKT	1	GSM3544603 [33]						
						Type																			
						Type																			
				Mucosal associated invariant T ce													MAIT	1	GSM3544603 [33]						
				Gamma Delta T cell (γδ T cells, γδ																					
					Gamma D												Gamma Delta 1 T ce	1	GSM3544603 [33]					[11]	
					Gamma D												Gamma Delta 2 T ce	1	GSM3544603 [33]					[11]	

*colored in light green means - important/main reference
* with background color light gray brown means it is book
*with background color blue means they have valuable data sets but not public

Reference	Area	
1	PBMC subset	Miltenyi Biotec (https://www.miltenyibiotec.com/US-en/resources/macs-handbook/human-cells-and-organs/human-cell-sources/blood-human.html)
2	DC	Starks, M. A. A. (2019). Immunology and Animal Biotechnology, EDTECH.
3	DC	Ouaguia, Laurissa, et al. "Circulating and hepatic BDCA1+, BDCA2+, and BDCA3+ dendritic cells are differentially subverted in patients with chronic HBV infection." Frontiers in immunology 10 (2019): 112.
4	DC	Tang-Huau, Tsing-Lee, et al. "Human in vivo-generated monocyte-derived dendritic cells and macrophages cross-present antigens through a vacuolar pathway." Nature communications 9.1 (2018): 1-12.
5	DC	R&D company file_Dendritic Cells - https://www.cell.com/pb-assets/products/nucleus/nucleus-phagocytes/rnd-systems-dendritic-cells-br.pdf
6	DC	Miltenyi Biotec Handbook https://www.miltenyibiotec.com/US-en/resources/macs-handbook/human-cells-and-organs/human-cell-types/dendritic-cells-human.html#structure-section-3d10a8d3-f4f1-49d9-b431-bcdfe0209a34
7	DC	Poltorak, Mateusz Pawel, and Barbara Ursula Schraml. "Fate mapping of dendritic cells." Frontiers in immunology 6 (2015): 199.
8	DC	Dress, Regine J., et al. "Plasmacytoid dendritic cells develop from Ly6D+ lymphoid progenitors distinct from the myeloid lineage." Nature immunology 20.7 (2019): 852-864.
9	B	Allman, David, and Shiv Pillai. "Peripheral B cell subsets." Current opinion in immunology 20.2 (2008): 149-157.
10	B	Wu, Yu-Chang Bryan, David Kipling, and Deborah K. Dunn-Walters. "The relationship between CD27 negative and positive B cell populations in human peripheral blood." Frontiers in immunology 2 (2011): 81.
11	PBMC subset	Ding, Yuan, et al. "Reference values for peripheral blood lymphocyte subsets of healthy children in China." Journal of Allergy and Clinical Immunology 142.3 (2018): 970-973.
12	PBMC, B, NK, frequency	Melzer, Susanne, et al. "Reference intervals for leukocyte subsets in adults: Results from a population-based study using 10-color flow cytometry." Cytometry Part B: Clinical Cytometry 88.4 (2015): 270-281.
13	B	LeBien, Tucker W., and Thomas F. Tedder. "B lymphocytes: how they develop and function." Blood 112.5 (2008): 1570-1580.
14	B	Marasco, Emiliano, et al. "B-cell activation with CD40L or CpG measures the function of B-cell subsets and identifies specific defects in immunodeficient patients." European journal of immunology 47.1 (2017): 131-143.
15	PBMC subset	Corkum, Christopher P., et al. "Immune cell subsets and their gene expression profiles from human PBMC isolated by Vacutainer Cell Preparation Tube (CPT™) and standard density gradient." BMC immunology 16.1 (2015): 1-18.
16	B	Piątosa, Barbara, et al. "B cell subsets in healthy children: reference values for evaluation of B cell maturation process in peripheral blood." Cytometry Part B: Clinical Cytometry 78.6 (2010): 372-381.
17	B	Morbach, H., et al. "Reference values for B cell subpopulations from infancy to adulthood." Clinical & Experimental Immunology 162.2 (2010): 271-279.
18	T, B, NK	Apoil, P. A., et al. "Reference values for T, B and NK human lymphocyte subpopulations in adults." Data in brief 12 (2017): 400-404.
19	T	Dekker, Linde, et al. "Reconstitution of t cell subsets following allogeneic hematopoietic cell transplantation." Cancers 12.7 (2020): 1974.
20	PBMC subset	Lepone, Lauren M., et al. "Analyses of 123 peripheral human immune cell subsets: defining differences with age and between healthy donors and cancer patients not detected in analysis of standard immune cell types." Journal of circulatir
21	Treg	Shevyrev, Daniil, and Valeriy Tereshchenko. "Treg Heterogeneity, Function, and Homeostasis." Frontiers in Immunology 10 (2019).
22	Treg	Mohr, Audrey, et al. "Human FOXP 3+ T regulatory cell heterogeneity." Clinical & Translational Immunology 7.1 (2018): e1005.
69	CD4 T cell	Fang, Difeng, and Jinfang Zhu. "Dynamic balance between master transcription factors determines the fates and functions of CD4 T cell and innate lymphoid cell subsets." Journal of Experimental Medicine 214.7 (2017): 1861-1876.
70	T cell classification	Zhuang, Quan, et al. "The detailed distribution of T cell subpopulations in immune-stable renal allograft recipients: a single center study." PeerJ 7 (2019): e6417.
71	CD8 T cell subset	van Aalderen, Michiel C., et al. "Label-free analysis of CD8+ T cell subset proteomes supports a progressive differentiation model of human-virus-specific T cells." Cell reports 19.5 (2017): 1068-1079.
72	CD4 T cell subset	Miltenyi Biotec (https://www.miltenyibiotec.com/US-en/resources/macs-handbook/human-cells-and-organs/human-cell-types/cd4-t-cells-human.html)
73	T memory cell	Gattinoni, Luca, et al. "T memory stem cells in health and disease." Nature medicine 23.1 (2017): 18-27.
74	CD4 T cell	Caccamo, Nadia, et al. "Atypical human effector/memory CD4+ T cells with a naive-like phenotype." Frontiers in immunology 9 (2018): 2832.
75	CD4 T cell subset, T subset	Golubovskaya, Vita, and Lijun Wu. "Different subsets of T cells, memory, effector functions, and CAR-T immunotherapy." Cancers 8.3 (2016): 36.
76	T cell development	Mockler, Mary B., Melissa J. Conroy, and Joanne Lysaght. "Targeting T cell immunometabolism for cancer immunotherapy; understanding the impact of the tumor microenvironment." Frontiers in oncology 4 (2014): 107.
77	T cell development	Benichou, Gilles, et al. "Role of memory T cells in allograft rejection and tolerance." Frontiers in immunology 8 (2017): 170.
78	T cell development	Opata, Michael M., et al. "Protection by and maintenance of CD4 effector memory and effector T cell subsets in persistent malaria infection." PLoS pathogens 14.4 (2018): e1006960.
79	T SCM	Restifo, Nicholas P. "Big bang theory of stem-like T cells confirmed." Blood 124.4 (2014): 476-477.
80	T cell development	Goulding, John, Vikas Tahiliani, and Shahram Salek-Ardakani. "OX40: OX40L axis: emerging targets for improving poxvirus-based CD8+ T-cell vaccines against respiratory viruses." Immunological reviews 244.1 (2011): 149-168.
81	CD4 T cell development	Pepper, Marion, and Marc K. Jenkins. "Origins of CD4+ effector and central memory T cells." Nature immunology 12.6 (2011): 467-471.
82	CD4 Treg	Simonetta, Federico, and Christine Bourgeois. "CD4+ FOXP3+ regulatory T-cell subsets in human immunodeficiency virus infection." Frontiers in immunology 4 (2013): 215.
83	PBMC subset	Kalina, Tomas, et al. "CD Maps–dynamic profiling of CD1 to CD100 surface expression on human leukocyte and lymphocyte subsets." Frontiers in immunology 10 (2019): 2434.
84	Circulating T memory cell	Farber, Donna L., Naomi A. Yudanin, and Nicholas P. Restifo. "Human memory T cells: generation, compartmentalization and homeostasis." Nature Reviews Immunology 14.1 (2014): 24-35.
85	effector T and memory T	Restifo NP, Gattinoni L. Lineage relationship of effector and memory T cells. Curr Opin Immunol. 2013;25(5):556-563. doi:10.1016/j.coi.2013.09.003
86	NK cell subset	Mahapatra, Sanjana, et al. "High-resolution phenotyping identifies NK cell subsets that distinguish healthy children from adults." PloS one 12.8 (2017): e0181134.
87	Treg frequency	Seddiki, Nabila, et al. "Persistence of naive CD45RA+ regulatory T cells in adult life." Blood 107.7 (2006): 2830-2838.
88	sification, immune signatures of CD4 TEMI	Tian, Yuan, et al. "Unique phenotypes and clonal expansions of human CD4 effector memory T cells re-expressing CD45RA." (2018): 51-12.
89	PBMC, T subsets frequency	Burel, Julie G., et al. "An integrated workflow to assess technical and biological variability of cell population frequencies in human peripheral blood by flow cytometry." The Journal of Immunology 198.4 (2017): 1748-1758.

Subset (abbreviation)	Anchor marker	CD subset measured
Total T cell	CD45	CD3 +/CD19 −
Helper T cell (CD4)	CD45	CD3 +/CD4 +
Cytotoxic T cell (CD8)	CD45	CD3 +/CD8 +
Total B cell	CD45	CD3 −/CD19 +
NK cell	CD45	CD3 −/CD16 +/CD56 +
TCRαβ + double-negative T (DNT) cell	CD3	CD3 +/CD4 −/CD8 −/TCRαβ +
γδT cell (γδT)	CD3	CD3 +/TCRγδ +
Double-positive T (DPT) cell	CD3	CD4 +/CD8 +
Naive helper T cell (CD4 Naive)	CD4	CD27 +/CD45RA +
Central memory helper T cell (CD4 CM)	CD4	CD27 +/CD45RA −
Effector memory helper T cell (CD4 EM)	CD4	CD27 −/CD45RA −
Terminally differentiated effector memory CD45RA + helper T cell (CD4 TEMRA)	CD4	CD45RA +/CD27 −
Naive cytotoxic T cell (CD8 Naive)	CD8	CD27 +/CD45RA +
Central memory cytotoxic T cell (CD8 CM)	CD8	CD27 +/CD45RA −
Effector memory cytotoxic T cell (CD8 EM)	CD8	CD27 −/CD45RA −
Terminally differentiated effector memory CD45RA + cytotoxic T cell (CD8 TEMRA)	CD8	CD45RA +/CD27 −
Naive B cell	CD19	CD27 −/IgD +
Memory B cell	CD19	CD27 +/IgD −
Transitional B cell	CD19	CD24 +/CD38 +
Plasmablasts	CD19	CD38 +/CD24 −

Panel	Population Name	Reliability	Corresponding Markers
T-cell	CD8 Activated	−	CD3+/CD8+/CD4− /CD38+/HLADR+
T-cell	CD4 Activated	+	CD3+/CD8−/CD4+/CD38+/HLADR+
T-cell	CD4 Central Memory	−	CD3+/CD8−/CD4+/CCR7+/CD45RA−
T-cell	CD8 Central Memory	−	CD3+/CD8+/CD4−/CCR7+/CD45RA−
T-cell	CD4 Effector	+	CD3+/CD8−/CD4+/CCR7−/CD45RA+
T-cell	CD8 Effector	+	CD3+/CD8+/CD4−/CCR7−/CD45RA+
T-cell	CD4 Effector Memory	+	CD3+/CD8−/CD4+/CCR7−/CD45RA−
T-cell	CD8 Effector Memory	−	CD3+/CD8+/CD4−/CCR7−/CD45RA−
T-cell	CD4 Naïve	+	CD3+/CD8−/CD4+/CCR7+/CD45RA+
T-cell	CD8 Naïve	+	CD3+/CD8+/CD4−/CCR7+/CD45RA+
B-cell	IgD−/CD27−	−	CD3−/CD19+/CD20+/IgD−/CD27−
B-cell	Transitional	+	CD3−/CD19+/CD20+
B-cell	Plasmablasts	−	CD3−/CD19+/CD20−/Cd24high/CD38high
B-cell	Naïve B	+	CD3−/CD19+/CD20+/CD27−/IgD+
B-cell	Memory IgD+	+	CD3−/CD19+/CD20+/IgD+/CD27+/IgD+
B-cell	CD19	+	CD3−/CD19+
B-cell	CD20	+	CD3−/CD20+
B-cell	Memory IgD-	+	CD3−/CD19+/CD20+/CD27+/IgD−
T-regulatory	Total T-regulatory	+	CD3+/CD4+/CD8-/LoCD127/HiCD25/CCR4+ (as % of CD4)
T-regulatory	Memory T-regulatory	+	CD3+/CD4+/CD8−/LoCD127/HiCD25/CCR4+/CD45RO+ (as % of total Treg)
T-regulatory	Naïve T-regulatory	+	CD3+/CD4+/CD8−/LoCD127/HiCD25/CCR4+/CD45RO− (as % of total Treg)
T-regulatory	CCR4-/CD45RO−	−	CD3+/CD4+/CD8−/LoCD127/HiCD25/CCR4−/CD45RO− (as % of parent)
DC/Mono/NK	CD14+/CD16+	−	CD14+/CD16+
DC/Mono/NK	CD16-/CD56+	+	CD16−/CD56+
DC/Mono/NK	CD16+/CD56-	−	CD16+/CD56−
DC/Mono/NK	CD16+/CD56+	+	CD16+/CD56+
DC/Mono/NK	HLADR+	−	HLADR+
DC/Mono/NK	Lin-CD14−	+	Lin-CD14−
DC/Mono/NK	Lin−/CD14+	+	Lin−/CD14+
DC/Mono/NK	CD16−/CD56−	−	CD16−/CD56−

Appendix 7 Supplemental Materials in Study III

- J. Zhong, M. Lyu, H. Jin, Z. Cao, L. T. Chitkushev, G. Zhang, D. B. Keskin, and V. Brusic, "Artificial Neural Networks for classification of single cell gene expression," bioRxiv, 2021.

❖ **Supplemental Table 1. Metadata describing samples as described by the sources.**

1. Datasets that are included in incremental learning experiments:

Source	Series	Date	Cell Type	Class	Sample Condition	Donors	Separation
10x Genomics	SRP073767	2017/01/16	CD19+ B cells	BC	healthy fresh blood	donor A, all cells	n/a
			CD14+ Monocytes	MC			
			CD56+ NK cells	NK			
			CD8+ CTLs (Cytotoxic T cells)	TC			
			CD4+CD45RO+ Memory T cells		healthy fresh PBMCs		
			CD4+CD25+ Treg cells				
			CD4+CD45RA+CD25- Naïve T cells				
			CD4+ Th cells				
			CD8+CD45RA+ Naïve CTLs (Cytotoxic T cells)				
GEO	GSM2773408	2018/10/15	CD14+CD16- Monocytes	MC	healthy fresh blood	donor 1	centrifuge, ficoll
	GSM2773409		CD14+CD16- Monocytes			donor 2	
	GSM3544603	2019/01/08	NK cells	NK	healthy fresh blood	two donors (IGPR998, IGPR999)	centrifuge, ficoll
			CD4+ T cells	TC			
			CD8+ T cells				
			iNKT (invariant Natural Killer T cells)				
			MAIT (Mucosal-associated Invariant T cells)				
			Gamma Delta 1 T cells				
			Gamma Delta 2 T cells				
	GSM3209407	2019/06/20	CD4+ T cells	TC	healthy fresh blood	one individual	centrifuge, ficoll
	GSM3209408		CD4+CCR5+CD69- T cells			one individual	
BroadS1	SCP345	2019/07	B cells	BC	healthy frozen blood	two donors (580/860), two ICA sites (Broad/Boston and MtSinai/NYC)	n/a
			Dendritic cells	DC			
			Monocytes	MC			
			NK cells	NK			
			T cells	TC			
BroadS2	GSE132044/ SCP424, SCP425 and SCP426	2020/04/06	B cells	BC	healthy frozen blood	pbmc1, pbmc2, different days	n/a
			CD4+ T cells	TC			
			CD14+ Monocytes	MC			
			CD16+ Monocytes				
			Cytotoxic T cells	TC			
			Dendritic cells	DC			
			NK cells	NK			
			Plasmacytoid Dendritic cells	DC			
			B cells	BC			
			CD4+ T cells	TC			
			CD14+ Monocytes	MC			
			CD16+ Monocytes				
			Cytotoxic T cells	TC			
			Dendritic cells	DC			
			NK cells	NK			
			Plasmacytoid Dendritic cells	DC			
			B cells	BC			
			CD4+ T cells	TC			
			CD14+ Monocytes	MC			
			CD16+ Monocytes				
			Cytotoxic T cells	TC			
			Dendritic cells	DC			
			NK cells	NK			
			B cells	BC			
			CD4+ T cells	TC			
			CD14+ Monocytes	MC			
			CD16+ Monocytes				
			Cytotoxic T cells	TC			
			Dendritic cells	DC			
			NK cells	NK			
			Plasmacytoid Dendritic cells	DC			

Sorting	Strategy	Other	Purity	Extraction & Sequencing	Reads	Upstream Alignment	Genome Build	Reference
bead-enriched from PBMC, negative selection	GemCode platform	Countess II Automated Cell Cou	~100% pure by FACS 98% pure by FACS 92% pure by FACS 98% pure by FACS 98% pure by FACS 95% pure by FACS 98% pure by FACS 99% pure by FACS 99% pure by FACS	Chromium Single-Cell 3' method (10X Genomics)	Illumina NextSeq500	10X Cell Ranger v1	GRCh37 (hg19)	Zheng et al, 2017
MACS	stained HLA-DR, CD14, CD16	magnetic beads	93-95% CD14+CD16-	Chromium Single-Cell 3' Reagent (v2) Kit (10X Genomics)	Illumina HiSeq 2500	10X Cell Ranger v1.3.1	GRCh37 (hg19)	Goudot et al, 2017
FACS	various	overnight fasting	n/a	Chromium Single-Cell 3' method (10X Genomics)	Illumina NextSeq 500	10X Cell Ranger v2.1.0	GRCh38 (hg38)	Gutierrez et al, 2019
FACS	Aria II (BD Biosciences)	n/a	n/a	Chromium Single-Cell 3' Reagent (v2) Kit (10X Genomics)	Illumina HiSeq 2500	10X Cell Ranger v2.0.1	GRCh38 (hg38)	Woodward et al, 2019
multiple regression and naive Bayes optimization, correlation	cells >=400 gene present	n/a	n/a	Chromium Single-Cell 3' method (10X Genomics)	n/a	10X Cell Ranger	n/a	n/a
Louvain community detection algorithm, k-NN, marker genes	pbmc1_10x_v2_A	n/a	n/a	Chromium Single-Cell 3' Reagent (v2) Kit (10X Genomics)	Illumina HiSeq 2500	10X Cell Ranger v1.2.0	GRCh38 (hg38)	Ding et al, 2020
	pbmc1_10x_v2_B			Chromium Single-Cell 3' Reagent (v2) Kit (10X Genomics)				
	pbmc1_10x_v3			Chromium Single-Cell 3' Reagent (v3) Kit (10X Genomics)				
	pbmc2_10X_v2			Chromium Single-Cell 3' Reagent (v2) Kit (10X Genomics)				

2. The non-representative datasets that are included in model vulnerability experiments:

Source	Series	Date	Cell Type	Class	Sample Condition
GEO	GSM3162632	2018/05/30	Tumor Ascites Dendritic cells	DC	tumor ascites
	GSM3162630		Tonsil Dendritic cells		tonsil tissue
	GSM3087629	2018/07/25	CD8+ T cells (methanol SSC)	TC	healthy frozen PBMCs
	GSM3430548	2018/11/07	IL-10 producing Foxp3-CD4+ T cells	TC	healthy blood
	GSM3430549		IL-10 producing Foxp3-CD4+ T cells		
	GSM3478792	2019/01/31	nonmalignant P5 CD3+CD5intSSCintCD4+ T cells	TC	patient fresh blood
	GSM3558027	2019/07/25	nonmalignant P5 CD3+CD5intSSCintCD4+ T cells (after therapy)		
	GSM3258345	2018/10/15	HLA-DR+ cells	MC	healthy fresh blood
	GSM3258347		HLA-DR+ cells (control)		
	GSM3258346		CD19+ cells	BC	
	GSM3258348		CD19+ cells (control)		
	GSM3087628	2018/07/25	CD8+ cells	TC	healthy fresh blood

Donors	Separation	Sorting	Strategy	Other
ovarian cancer patients	centrifuge, ficoll	bead-enriched, negative selec	gated as HLA-DR+CD11c+CD1c+CD16−	cell culture >10 days
healthy patients (both male and female)			gated as HLA-DR+CD11c+CD14−	
anonymous, healthy donors from NIH Blood Bank	LeucoSep tube w	MACS	Dynabeads™ CD8 Positive Isolation Kit	methanol fixation
healthy donor 1	Biocoll separatio	MACS	enriched using MACS CD4 beads (Miltenyi)	activated cells
healthy donor 2				
61-year-old male patient donor, with stage IVA Sézary s	centrifuge, ficoll	FACS	Aria II (BD Biosciences)	activated cells
healthy donor	centrifuge, ficoll	FACS	designed to target live HLA-DR+ cells and de	enriched, mixed populations
control				
healthy donor			designed to target live CD19+ cells and deple	
control				
n/a	centrifuge, ficoll	MACS	Dynabeads™ CD8+ Isolation Kit	magnetic beads

Purity	Extraction & Sequencing	Reads	Upstream Alignment	Genome Build	Reference
n/a	Chromium Single-Cell 3′ Reagent (v2) Kit (10X Genomics)	Illumina HiSeq 2500	10X Cell Ranger v2.0.1	GRCh38 (hg38)	Tang-Huau et al, 2018
n/a	Chromium Single-Cell 3′ Reagent (v2) Kit (10X Genomics)	Illumina NextSeq 500	10X Cell Ranger v2.0.1	GRCh38 (hg38)	Chen et al, 2018
n/a	Chromium Single-Cell 3' method (10X Genomics)	Illumina HiSeq 4000	10X Cell Ranger	GRCh37 (hg19)	Brockmann et al, 2018
n/a	Chromium Single-Cell 5' method (10X Genomics)	Illumina HiSeq 4000	10X Cell Ranger v2.2	GRCh38 (hg38)	Borcherding et al, 2019
			10X Cell Ranger v2.1		
n/a	Chromium Single-Cell 3' method (10X Genomics)	Illumina MiSeq	10X Cell Ranger v1.3.1	GRCh37 (hg19)	Ranu et al, 2019
		HiSeq X Ten			
		Illumina MiSeq			
		HiSeq X Ten			
n/a	Chromium Single-Cell 3′ Reagent (v2) Kit (10X Genomics)	Illumina HiSeq 3000	10X Cell Ranger v2.0.1	GRCh38 (hg38)	Chen et al, 2018

❖ Supplemental Table 2. The results of basic statistical analysis of the data sets.

LEGEND

Q1,Q2,Q3,Q4: Quartiles

IQR: InterQuartile Range, Q3-Q1

R=Range=Max-Min

Below QC threshold (670-300)

Above QC threshold (670-300)

1. Data sets that are included in incremental learning experiments:

Source	Series	Date	Cell Type	Class	Strategy	Tag	Cell Number (N)
10x Genomics	SRP073767	2017/01/16	CD19+ B cells	BC	GemCode platform	BC01	10085
			CD14+ Monocytes	MC		MC01	2612
			CD56+ NK cells	NK		NK01	8385
			CD8+ CTLs (Cytotoxic T cells)	TC		TC01	10209
			CD4+CD45RO+ Memory T cells			TC02	10224
			CD4+CD25+ Treg cells			TC03	10263
			CD4+CD45RA+CD25- Naive T cells			TC04	10479
			CD4+ Th cells			TC05	11213
			CD8+CD45RA+ Naive CTLs (Cytotoxic T cells)			TC06	11953
GEO	GSM2773408	2018/10/15	CD14+CD16- Monocytes	MC	stained HLA-DR, CD14, CD16	MC02	425
	GSM2773409		CD14+CD16- Monocytes			MC03	431
	GSM3544603	2019/01/08	NK cells	NK	various	NK02	309
			CD4+ T cells	TC		TC07	222
			CD8+ T cells			TC08	310
			iNKT (invariant Natural Killer T cells)			TC09	325
			MAIT (Mucosal-associated Invariant T cells)			TC10	382
			Gamma Delta 1 T cells			TC11	284
			Gamma Delta 2 T cells			TC12	204
	GSM3209407	2019/06/20	CD4+ T cells	TC	Aria II (BD Biosciences)	TC13	965
	GSM3209408		CD4+CCR5+CD69- T cells			TC14	435
BroadS1	SCP345	2019/07	B cells	BC	cells >=400 gene present	BC02	1660
			Dendritic cells	DC		DC01	142
			Monocytes	MC		MC04	1661
			NK cells	NK		NK03	1394
			T cells	TC		TC15	8326
BroadS2	GSE132044/ SCP424, SCP425 and SCP426	2020/04/06	B cells	BC	pbmc1_10x_v2_A	BC03	288
			CD4+ T cells	TC		TC16	550
			CD14+ Monocytes	MC		MC05	640
			CD16+ Monocytes			MC06	102
			Cytotoxic T cells	TC		TC17	1174
			Dendritic cells	DC		DC02	55
			NK cells	NK		NK04	166
			Plasmacytoid Dendritic cells	DC		DC03	26
			B cells	BC	pbmc1_10x_v2_B	BC04	388
			CD4+ T cells	TC		TC18	908
			CD14+ Monocytes	MC		MC07	379
			CD16+ Monocytes			MC08	73
			Cytotoxic T cells	TC		TC19	954
			Dendritic cells	DC		DC04	33
			NK cells	NK		NK05	263
			Plasmacytoid Dendritic cells	DC		DC05	12
			B cells	BC	pbmc1_10x_v3	BC05	346
			CD4+ T cells	TC		TC20	960
			CD14+ Monocytes	MC		MC09	354
			CD16+ Monocytes			MC10	98
			Cytotoxic T cells	TC		TC21	962
			Dendritic cells	DC		DC06	38
			NK cells	NK		NK06	194
			B cells	BC	pbmc2_10X_v2	BC06	862
			CD4+ T cells	TC		TC22	962
			CD14+ Monocytes	MC		MC11	436
			CD16+ Monocytes			MC12	50
			Cytotoxic T cells	TC		TC23	694
			Dendritic cells	DC		DC07	76
			NK cells	NK		NK07	219
			Plasmacytoid Dendritic cells	DC		DC08	30

Column_Sum (total number of counts in each cell)											
Min	% <670	Q1 (25%)	Q2/ Median (50%)	Mean	Q3 (75%)	Max	Range (R)	IQR	Standard Deviation (σ)	Skewness (Sk)	Kurtosis (K)
460	3.24	1029	1231	1424	1601	6862	6402	572	663.68	2.07	5.85
567	29.20	653	776	1079	1034	9005	8438	381	966.31	3.86	16.11
445	2.42	1349	1576	1661	1850	6451	6006	501	605.60	1.59	5.29
432	1.05	1344	1620	1671	1901	7632	7200	557	543.90	1.46	5.79
496	0.94	1238	1500	1609	1835	10255	9759	597	601.76	2.50	17.14
383	3.88	1010	1218	1301	1475	7278	6895	465	490.82	2.37	12.86
358	3.90	980	1181	1222	1385	6775	6417	405	395.98	1.91	10.28
416	3.26	1076	1320	1390	1585	8767	8351	509	520.59	2.37	15.95
334	0.84	1229	1435	1505	1673	4991	4657	444	462.11	1.46	3.91
853	0	2252	3303	3641	4661	11194	10341	2409	1812.52	0.80	0.68
871	0	2174	3282	3574	4530	15009	14138	2355	2005.85	1.62	5.08
2627	0	2965	3107	3113	3263	3644	1017	298	206.52	0.21	-0.46
2191	0	2664	2745	2791	2909	3291	1100	244	191.17	0.37	0.45
2505	0	2698	2803	2865	3045	3369	864	347	210.10	0.55	-0.81
2177	0	2764	2952	2949	3149	3521	1344	385	248.10	-0.34	-0.13
2206	0	2751	2827	2903	3105	3445	1239	353	222.47	0.30	-0.35
2167	0	2817	3033	3018	3198	3541	1374	381	232.98	-0.18	-0.52
2544	0	2773	2944	2966	3159	3522	978	385	224.20	0.12	-1.13
447	9.64	2360	2844	2755	3399	9134	8687	1039	1157.46	-0.03	1.48
417	5.52	2462	2793	2771	3193	9134	8717	731	955.77	0.53	6.25
2026	0	2551	2815	3069	3355	7227	5201	804	775.50	1.82	3.92
2614	0	4386	4880	4860	5327	7106	4492	941	721.39	-0.02	0.70
2146	0	2664	3040	3295	3945	5691	3545	1281	801.33	0.71	-0.53
2148	0	2969	3210	3276	3489	5884	3736	520	484.70	1.27	3.41
1845	0	2818	3138	3179	3480	7782	5937	662	560.48	1.19	4.71
626	0.69	1033	1190	1192	1326	1969	1343	293	217.98	0.42	0.58
724	0	1130	1293	1287	1410	2594	1870	279	226.35	0.91	3.44
366	6.56	962	1171	1181	1405	1972	1606	443	332.75	0.03	-0.30
824	0	1210	1476	1466	1660	2223	1399	449	313.80	0.25	-0.51
672	0	1123	1251	1254	1385	2173	1501	262	198.75	0.20	0.58
946	0	1110	1320	1388	1585	2100	1154	475	331.44	0.59	-0.69
955	0	1307	1400	1409	1522	1938	983	215	185.01	0.19	0.20
913	0	1058	1455	1454	1860	2007	1094	802	405.35	0.06	-1.68
717	0	1057	1168	1189	1278	2052	1335	221	211.11	1.08	2.13
361	0.33	1196	1291	1294	1388	2117	1756	192	181.73	0.20	3.74
270	5.54	1063	1229	1207	1386	2018	1748	323	292.74	-0.49	0.80
946	0	1479	1652	1620	1843	2100	1154	364	307.64	-0.60	-0.30
455	0.10	1150	1258	1263	1355	1971	1516	204	157.39	0.54	1.80
971	0	1054	1198	1296	1301	2040	1069	247	325.02	1.21	0.29
999	0	1206	1310	1343	1439	2127	1128	233	196.08	1.09	2.03
947	0	1079	1567	1456	1713	1970	1023	634	380.56	-0.06	-1.72
720	0	12956	1520	1419	1655	2155	1435	359	324.59	-0.78	-0.45
709	0	937	1599	1436	1790	2238	1529	853	425.39	-0.35	-1.43
798	0	1155	1367	1498	1884	2420	1622	729	411.23	0.55	-0.95
993	0	1203	1300	1339	1445	2321	1328	242	198.89	1.76	6.43
774	0	1610	1737	1726	1866	2249	1475	256	208.88	-0.77	1.98
873	0	1071	1164	1172	1240	1527	654	169	156.84	0.42	0.22
1000	0	1707	1843	1803	1967	2350	1350	260	267.41	-1.03	1.24
631	0.12	1260	1402	1397	1526	2478	1847	266	228.63	0.24	1.30
85	0.73	1272	1432	1403	1542	2302	2217	270	242.42	-0.99	5.23
521	0.92	1237	1533	1503	1723	2385	1864	486	345.50	-0.05	-0.37
979	0	1055	1415	1553	2068	2372	1393	1013	508.80	0.25	-1.68
810	0	1391	1521	1505	1622	2324	1514	231	206.84	-0.07	1.73
903	0	1149	1253	1311	1371	2141	1238	222	268.94	1.38	1.82
942	0	1362	1511	1526	1667	2329	1387	305	227.74	0.49	1.29
926	0	1030	1163	1423	1879	2364	1438	849	522.83	0.88	-0.97

Column_Positive (number of genes with counts > 0)												Reference
Min	%<300	Q1 (25%)	Q2/ Median (50%)	Mean	Q3 (75%)	Max	Range (R)	IQR	Standard Deviation (σ)	Skewness (Sk)	Kurtosis (K)	
197	2.82	417	474	524	574	1854	1657	157	178.62	1.87	4.71	
267	4.10	332	378	455	461	2393	2126	129	244.58	3.37	12.61	
223	0.80	614	701	722	798	2073	1850	184	200.08	1.18	3.68	
213	0.91	501	567	578	633	2216	2003	132	140.12	1.51	7.77	
217	0.81	482	552	579	635	2677	2460	153	164.45	2.44	15.23	Zheng et al, 2017
190	1.98	463	542	562	625	2311	2121	162	165.61	2.04	10.05	
188	2.73	421	486	496	550	2130	1942	129	122.74	1.60	8.45	
204	1.94	460	541	558	626	2435	2231	166	162.65	1.99	11.53	
141	1.53	442	496	511	556	1348	1207	114	117.10	1.21	3.17	
382	0	907	1186	1217	1496	2715	2333	589	423.06	0.38	-0.09	Goudot et al, 2017
315	0	876	1163	1184	1425	3402	3087	548	437.55	0.90	1.99	
557	0	788	864	878	964	1975	1418	176	154.83	1.45	7.78	
487	0	774	883	870	966	1417	930	191	168.05	-0.06	0.46	
492	0	853	930	932	1016	1360	868	163	153.78	-0.21	0.72	
514	0	807	890	903	985	1729	1215	178	176.91	1.19	4.36	Gutierrez et al, 2019
503	0	849	942	922	1018	1484	981	169	156.62	-0.42	1.04	
461	0	832	928	922	1015	1563	1102	183	164.87	-0.01	1.56	
500	0	834	919	915	998	1691	1191	164	178.36	0.71	3.55	
32	4.87	834	955	899	1057	2548	2516	223	293.06	-0.46	2.13	Woodward et al, 2019
56	2.53	889	988	960	1071	2548	2492	183	262.07	-0.14	5.80	
489	0	697	790	952	986	4286	3797	289	490.49	2.99	10.54	
695	0	1661	1890	1875	2096	2986	2291	435	397.53	-0.17	0.77	
490	0	653	839	938	1170	2351	1861	517	353.18	0.91	0.13	n/a
489	0	802	902	920	1001	2365	1876	199	194.21	1.85	8.87	
486	0	797	907	935	1029	4368	3882	232	240.52	2.86	21.29	
230	1.74	582	739	770	949	1648	1418	367	265.63	0.46	-0.02	
382	0	678	884	878	1056	2252	1870	378	262.97	0.65	1.63	
89	13.28	410	611	679	913	1658	1569	503	360.45	0.64	-0.36	
355	0	731	988	1019	1241	1859	1504	510	347.32	0.46	-0.62	
290	0.09	555	775	763	946	1859	1569	391	242.05	0.35	0.08	
593	0	749	981	1024	1228	1736	1143	478	322.59	0.67	-0.50	
452	0	757	952	931	1065	1631	1179	308	219.97	0.38	0.32	
591	0	718	912	1064	1525	1662	1071	806	413.32	0.38	-1.66	
358	0	680	784	814	893	1728	1370	213	223.93	1.15	2.36	
88	0.33	805	911	914	1016	1749	1661	211	195.76	0.34	2.38	
54	7.65	512	676	707	901	1686	1632	389	301.99	0.38	0.13	
563	0	1083	1256	1231	1465	1772	1209	382	318.55	-0.48	-0.49	
136	0.10	769	862	862	943	1636	1500	174	152.95	0.60	3.35	
570	0	700	840	933	944	1670	1100	244	318.17	1.18	0.29	
489	0	818	913	934	1010	1755	1266	192	198.76	1.22	3.02	
576	0	723	1199	1104	1380	1637	1061	657	386.95	-0.04	-1.69	Ding et al, 2020
402	0	970	1184	1092	1323	1817	1415	353	323.38	-0.73	-0.49	
391	0	614	1268	1112	1465	1915	1524	851	423.13	-0.34	-1.44	
350	0	821	1034	1166	1534	2119	1769	713	417.50	0.51	-0.92	
682	0	873	977	1008	1115	1980	1298	242	196.79	1.75	6.10	
485	0	1283	1405	1396	1534	1935	1450	250	211.80	-0.85	2.23	
570	0	743	840	845	909	1321	751	166	162.46	0.91	1.32	
664	0	1378	1508	1473	1618	2037	1373	240	260.52	-0.93	1.18	
201	0.23	849	1003	997	1165	2185	1984	316	273.51	0.15	0.66	
14	0.83	857	1062	1028	1209	1970	1956	352	271.67	-0.44	0.93	
164	2.52	768	1046	1058	1342	2041	1877	574	401.39	0.03	-0.65	
598	0	708	963	1183	1679	2030	1432	971	508.93	0.31	-1.65	
346	0	972	1120	1083	1227	2019	1673	255	254.89	-0.22	1.11	
577	0	815	914	977	1040	1794	1217	225	265.69	1.31	1.58	
420	0	938	1107	1070	1229	1993	1573	295	270.23	0.06	1.31	
581	0	685	857	1083	1494	2049	1468	808	523.84	0.93	-0.85	

2. The non-representative data sets that are included in model vulnerability experiments:

Source	Series	Date	Cell Type	Class	Strategy	Group	Cell Number (N)
GEO	GSM3162632	2018/05/30	Tumor Ascites Dendritic cells	DC	tumor tissue	Other Tissue	1613
	GSM3162630		Tonsil Dendritic cells		tonsil tissue		2739
	GSM3087629	2018/07/25	CD8+ T cells (methanol SSC)	TC	methanol fixation	Dead Cells	4753
	GSM3430548	2018/11/07	IL-10 producing Foxp3-CD4+ T cells	TC	IL-10 producing	Activated Cells	1247
	GSM3430549		IL-10-producing Foxp3-CD4+ T cells				1902
	GSM3478792	2019/01/31	nonmalignant P5 CD3+CD5intSSCintCD4+ T cells	TC	functional study	Activated Cells	4486
	GSM3558027	2019/07/25	nonmalignant P5 CD3+CD5intSSCintCD4+ T cells (after therapy)				3725
	GSM3258345	2018/10/15	HLA-DR+ cells	MC	selected by designed panel	Mixed Population	48
	GSM3258347		HLA-DR+ cells (control)				2397
	GSM3258346		CD19+ cells	BC			26
	GSM3258348		CD19+ cells (control)				1760
	GSM3087628	2018/07/25	CD8+ cells	TC	selected by designed panel	Mixed Population	5662

Column_Sum (total number of counts in each cell)											
Min	% <670	Q1 (25%)	Q2/ Median (50%)	Mean	Q3 (75%)	Max	Range (R)	IQR	Standard Deviation (σ)	Skewness (Sk)	Kurtosis (K)
675	0	2122	3004	3080	3877	11511	10836	1755	1357.05	1.16	3.98
825	0	5323	7081	9119	10309	62353	61528	4987	6397.48	2.90	11.98
835	0	1787	2686	2790	3402	33385	32550	1615	1531.10	4.53	60.15
1424	0	4341	5855	6345	7860	25281	23857	3520	2835.92	1.37	4.02
815	0	2733	3631	3893	4832	16781	15966	2099	1732.94	1.28	4.44
1575	0	4017	4969	5158	5910	27095	25520	1893	2100.48	2.30	12.08
1058	0	3872	4615	4797	5413	29910	28852	1541	1663.79	3.02	27.48
421	2.08	3795	6270	7039	9530	18584	18163	5735	4119.06	0.80	0.26
1058	0	2240	3316	3771	4724	21431	20373	2484	2190.60	2.02	7.21
22	7.69	2673	4288	4320	5797	8445	8423	3124	2250.02	-0.23	-0.61
1951	0	2972	4067	5252	5679	50189	48238	2707	4212.31	3.97	22.74
980	0	2924	3455	3681	4145	57391	56411	1221	1533.98	9.14	273.39

Column_Positive (number of genes with counts > 0)												Reference
Min	%<300	Q1 (25%)	Q2/ Median (50%)	Mean	Q3 (75%)	Max	Range (R)	IQR	Standard Deviation (σ)	Skewness (Sk)	Kurtosis (K)	
218	0.81	797	965	959	1110	2695	2477	313	292.71	0.90	4.43	Tang-Huau et al, 2018
401	0	1526	1848	2089	2397	6354	5953	872	829.60	1.55	2.61	
309	0	612	815	814	959	4369	4060	347	284.57	1.89	12.34	Chen et al, 2018
479	0	1589	2031	2047	2511	4638	4159	922	671.89	0.30	0.15	Brockmann et al, 2018
311	0	875	1162	1203	1458	3705	3394	583	467.31	0.83	1.63	
94	0.02	1246	1458	1500	1690	5147	5053	444	471.00	1.37	5.01	Borcherding et al, 2019
60	0.08	1117	1268	1310	1467	4859	4799	350	336.16	1.76	10.53	
233	2.08	1181	1474	1477	1839	2751	2518	658	522.09	-0.16	0.02	Ranu et al, 2019
38	1.75	903	1205	1239	1541	3911	3873	638	481.48	0.61	1.41	
20	30.77	236	747	751	1199	1546	1526	963	494.13	-0.07	-1.48	
78	1.25	974	1228	1402	1583	5285	5207	609	665.77	2.04	5.98	
336	0	869	963	998	1075	5717	5381	206	258.82	2.92	28.55	Chen et al, 2018

❖ Supplemental Table 3. The assessment of classification performance for incremental learning by cycles and steps.

LEGEND

2-fold cross-validation
New set classification
Test Result (BroadS1)
Final Result - BroadS1
Final Result - BroadS2
Nan - not analyzed

				Data Sets	TP	TN	FP	FN	# of Cells	ACC	SE	SP	PR	RE	F1	ACC
Cycle 0	Step 1	2-fold cross validation	B cells	BC01	10078	75330	8	7	85423	0.99982	0.99931	0.99989	0.99921	0.99931	0.99926	0.99865
			Monocytes	MC01	2582	82780	31	30	85423	0.99929	0.98851	0.99963	0.98814	0.98851	0.98832	
			NK cells	NK01	8358	77016	22	27	85423	0.99943	0.99678	0.99971	0.99737	0.99678	0.99707	
			T cells	TC01-TC06	64290	21028	54	51	85423	0.99877	0.99921	0.99744	0.99916	0.99921	0.99918	
	Step 2	added-predict	Monocytes	MC02	374	0	0	51	425	0.88000	0.88000	NA	1.00000	0.88000	0.93617	0.82009
	Step 3	added-predict	Monocytes	MC03	328	0	0	103	431	0.76102	0.76102	NA	1.00000	0.76102	0.86429	
	Step 4	BroadS1-test	B cells	BC02	1378	11523	0	282	13183	0.97861	0.83012	1.00000	1.00000	0.83012	0.90718	0.81863
			Monocytes	MC04	1483	11403	119	178	13183	0.97747	0.89284	0.98967	0.92572	0.89284	0.90898	
			Dendritic cells	DC01	0	0	0	142	13183	0.00000	0.00000	NA	NA	0.00000	0.00000	
			NK cells	NK03	1377	9546	2243	17	13183	0.82857	0.98780	0.80974	0.38039	0.98780	0.54926	
			T cells	TC15	6554	4828	29	1772	13183	0.86338	0.78717	0.99403	0.99559	0.78717	0.87920	
Cycle 1	Step 5	2-fold cross validation	B cells	BC01	10074	76187	7	11	86279	0.99979	0.99891	0.99991	0.99931	0.99891	0.99911	0.99842
			Monocytes	MC01-MC03	3436	82770	41	32	86279	0.99915	0.99077	0.99950	0.98821	0.99077	0.98949	
			NK cells	NK01	8341	77881	13	44	86279	0.99934	0.99475	0.99983	0.99844	0.99475	0.99659	
			T cells	TC01-TC06	64292	21863	75	49	86279	0.99856	0.99924	0.99658	0.99883	0.99924	0.99903	
	Step 6	added-predict	NK cells	NK02	309	0	0	0	309	1.00000	1.00000	NA	1.00000	1.00000	1.00000	0.24263
	Step 7	added-predict	T cells	TC07	56	0	0	166	222	0.25225	0.25225	NA	1.00000	0.25225	0.40287	
	Step 8	added-predict	T cells	TC08	97	0	0	213	310	0.31290	0.31290	NA	1.00000	0.31290	0.47665	
	Step 9	added-predict	T cells	TC09	6	0	0	319	325	0.01846	0.01846	NA	1.00000	0.01846	0.03625	
	Step 10	added-predict	T cells	TC10	7	0	0	375	382	0.01832	0.01832	NA	1.00000	0.01832	0.03598	
	Step 11	added-predict	T cells	TC11	10	0	0	274	284	0.03521	0.03521	NA	1.00000	0.03521	0.06802	
	Step 12	added-predict	T cells	TC12	9	0	0	195	204	0.04412	0.04412	NA	1.00000	0.04412	0.08451	
	Step 13	BroadS1-test	B cells	BC02	1159	11523	0	501	13183	0.96200	0.69819	1.00000	1.00000	0.69819	0.82228	0.78230
			Monocytes	MC04	1661	10912	610	0	13183	0.95373	1.00000	0.94706	0.73140	1.00000	0.84487	
			Dendritic cells	DC01	0	0	0	142	13183	0.00000	0.00000	NA	NA	0.00000	0.00000	
			NK cells	NK03	1371	9572	2217	23	13183	0.83008	0.98350	0.81194	0.38211	0.98350	0.55038	
			T cells	TC15	6122	4814	43	2204	13183	0.82955	0.73529	0.99115	0.99303	0.73529	0.84494	
Cycle 2	Step 14	2-fold cross validation	B cells	BC01	10080	78219	11	5	88315	0.99982	0.99950	0.99986	0.99891	0.99950	0.99920	0.99808
			Monocytes	MC01-MC03	3406	84825	22	62	88315	0.99905	0.98212	0.99974	0.99358	0.98212	0.98782	
			NK cells	NK01-NK02	8634	79594	27	60	88315	0.99901	0.99310	0.99966	0.99688	0.99310	0.99499	
			T cells	TC01-TC12	66025	22137	110	43	88315	0.99827	0.99935	0.99506	0.99834	0.99935	0.99884	
	Step 15	added-predict	T cells	TC13	956	0	0	9	965	0.99067	0.99067	NA	1.00000	0.99067	0.99531	0.99143
	Step 16	added-predict	T cells	TC14	432	0	0	3	435	0.99310	0.99310	NA	1.00000	0.99310	0.99654	
	Step 17	BroadS1-test	B cells	BC02	1431	11523	0	229	13183	0.98263	0.86205	1.00000	1.00000	0.86205	0.92591	0.92217
			Monocytes	MC04	1624	11361	161	37	13183	0.98498	0.97772	0.98603	0.90980	0.97772	0.94254	
			Dendritic cells	DC01	0	0	0	142	13183	0.00000	0.00000	NA	NA	0.00000	0.00000	
			NK cells	NK03	931	11616	173	463	13183	0.95176	0.66786	0.98533	0.84330	0.66786	0.74540	
			T cells	TC15	8171	4165	692	155	13183	0.93575	0.98138	0.85753	0.92192	0.98138	0.95072	
Cycle 3	Step 18	2-fold cross validation	B cells	BC01	10081	79615	15	4	89715	0.99979	0.99960	0.99981	0.99851	0.99960	0.99905	0.99819
			Monocytes	MC01-MC03	3411	86226	21	57	89715	0.99913	0.98356	0.99976	0.99388	0.98356	0.98869	
			NK cells	NK01-NK02	8642	80991	30	52	89715	0.99909	0.99402	0.99963	0.99654	0.99402	0.99528	
			T cells	TC01-TC14	67419	22151	96	49	89715	0.99838	0.99927	0.99568	0.99858	0.99927	0.99892	
	Step 19	added-predict	B cells	BC03	240	0	0	48	288	0.83333	0.83333	NA	1.00000	0.83333	0.90909	0.91869
	Step 20	added-predict	T cells	TC16	539	0	0	11	550	0.98000	0.98000	NA	1.00000	0.98000	0.98990	
	Step 21	added-predict	Monocytes	MC05	640	0	0	0	640	1.00000	1.00000	NA	1.00000	1.00000	1.00000	
	Step 22	added-predict	Monocytes	MC06	102	0	0	0	102	1.00000	1.00000	NA	1.00000	1.00000	1.00000	
	Step 23	added-predict	T cells	TC17	1108	0	0	66	1174	0.94378	0.94378	NA	1.00000	0.94378	0.97108	
	Step 24	added-predict	Dendritic cells	DC02	0	0	0	55	55	0.00000	0.00000	NA	0.00000	0.00000	0.00000	
	Step 25	added-predict	NK cells	NK04	128	0	0	38	166	0.77108	0.77108	NA	1.00000	0.77108	0.87075	
	Step 26	added-predict	pDC	DC03	0	0	0	26	26	0.00000	0.00000	NA	0.00000	0.00000	0.00000	
	Step 27	BroadS1-test	B cells	BC02	1444	11523	0	216	13183	0.98362	0.86988	1.00000	1.00000	0.86988	0.93041	0.92953
			Monocytes	MC04	1652	11344	178	9	13183	0.98582	0.99458	0.98455	0.90273	0.99458	0.94643	
			Dendritic cells	DC01	0	0	0	142	13183	0.00000	0.00000	NA	NA	0.00000	0.00000	
			NK cells	NK03	1058	11529	260	336	13183	0.95479	0.75897	0.97795	0.80273	0.75897	0.78024	
			T cells	TC15	8100	4366	491	226	13183	0.94561	0.97286	0.89891	0.94285	0.97286	0.95762	

Cycle 4	Step 28	2-fold cross validation	B cells	BC01, BC03	10364	82308	35	9	92716	0.99953	0.99913	0.99957	0.99663	0.99913	0.99788	0.99612
			Monocytes	MC01-MC03, MC05-MC06	4150	88435	71	60	92716	0.99859	0.98575	0.99920	0.98318	0.98575	0.98446	
			Dendritic cells	DC02-DC03	0	92635	0	81	92716	0.99913	0.00000	1.00000	NA	0.00000	0.00000	
			NK cells	NK01-NK02, NK04	8724	83795	61	136	92716	0.99788	0.98465	0.99927	0.99306	0.98465	0.98884	
			T cells	TC01-TC14, TC16-TC17	69118	23331	193	74	92716	0.99712	0.99893	0.99180	0.99722	0.99893	0.99807	
	Step 29	added-predict	B cells	BC04	377	0	0	11	388	0.97165	0.97165	NA	1.00000	0.97165	0.98562	0.93721
	Step 30	added-predict	T cells	TC18	903	0	0	5	908	0.99449	0.99449	NA	1.00000	0.99449	0.99724	
	Step 31	added-predict	Monocytes	MC07	378	0	0	1	379	0.99736	0.99736	NA	1.00000	0.99736	0.99868	
	Step 32	added-predict	Monocytes	MC08	73	0	0	0	73	1.00000	1.00000	NA	1.00000	1.00000	1.00000	
	Step 33	added-predict	T cells	TC19	942	0	0	12	954	0.98742	0.98742	NA	1.00000	0.98742	0.99367	
	Step 34	added-predict	Dendritic cells	DC04	24	0	0	9	33	0.72727	0.72727	NA	1.00000	0.72727	0.84210	
	Step 35	added-predict	NK cells	NK05	113	0	0	150	263	0.42966	0.42966	NA	1.00000	0.42966	0.60107	
	Step 36	added-predict	pDC	DC05	11	0	0	1	12	0.91667	0.91667	NA	1.00000	0.91667	0.95652	
	Step 37	BroadS1-test	B cells	BC02	1501	11521	2	159	13183	0.98779	0.90422	0.99983	0.99867	0.90422	0.94910	0.93120
			Monocytes	MC04	1637	11389	133	24	13183	0.98809	0.98555	0.98846	0.92486	0.98555	0.95424	
			Dendritic cells	DC01	90	13022	19	52	13183	0.99461	0.63380	0.99854	0.82569	0.63380	0.71713	
			NK cells	NK03	853	11659	130	541	13183	0.94910	0.61191	0.98897	0.86775	0.61191	0.71771	
			T cells	TC15	8195	4234	623	131	13183	0.94281	0.98427	0.87173	0.92935	0.98427	0.95602	
Cycle 5	Step 38	2-fold cross validation	B cells	BC01, BC03-BC04	10744	84952	13	17	95726	0.99969	0.99842	0.99985	0.99879	0.99842	0.99860	0.99540
			Monocytes	MC01-MC03, MC05-MC08	4607	90982	82	55	95726	0.99857	0.98820	0.99910	0.98251	0.98820	0.98535	
			Dendritic cells	DC02-DC05	70	95598	2	56	95726	0.99939	0.55556	0.99998	0.97222	0.55556	0.70707	
			NK cells	NK01-NK02, NK04-NK05	8908	86534	69	215	95726	0.99703	0.97643	0.99920	0.99231	0.97643	0.98431	
			T cells	TC01-TC14, TC16-TC19	70957	24398	274	97	95726	0.99612	0.99863	0.98889	0.99615	0.99863	0.99739	
	Step 39	added-predict	B cells	BC05	344	0	0	2	346	0.99422	0.99422	NA	1.00000	0.99422	0.99710	0.96917
	Step 40	added-predict	T cells	TC20	946	0	0	14	960	0.98542	0.98542	NA	1.00000	0.98542	0.99266	
	Step 41	added-predict	Monocytes	MC09	353	0	0	1	354	0.99718	0.99718	NA	1.00000	0.99718	0.99859	
	Step 42	added-predict	Monocytes	MC10	98	0	0	0	98	1.00000	1.00000	NA	1.00000	1.00000	1.00000	
	Step 43	added-predict	T cells	TC21	938	0	0	24	962	0.97505	0.97505	NA	1.00000	0.97505	0.98737	
	Step 44	added-predict	Dendritic cells	DC06	30	0	0	8	38	0.78947	0.78947	NA	1.00000	0.78947	0.88235	
	Step 45	added-predict	NK cells	NK06	152	0	0	42	194	0.78351	0.78351	NA	1.00000	0.78351	0.87862	
	Step 46	BroadS1-test	B cells	BC02	1526	11517	6	134	13183	0.98938	0.91928	0.99948	0.99608	0.91928	0.95614	0.93545
			Monocytes	MC04	1610	11457	65	51	13183	0.99120	0.96930	0.99436	0.96119	0.96930	0.96523	
			Dendritic cells	DC01	6	13032	9	136	13183	0.98900	0.04225	0.99931	0.40000	0.04225	0.07643	
			NK cells	NK03	1083	11436	353	311	13183	0.94963	0.77690	0.97006	0.75418	0.77690	0.76537	
			T cells	TC15	8107	4439	418	219	13183	0.95168	0.97370	0.91394	0.95097	0.97370	0.96220	
Cycle 6	Step 47	2-fold cross validation	B cells	BC01, BC03-BC05	11090	87551	20	17	98678	0.99963	0.99847	0.99977	0.99820	0.99847	0.99833	0.99435
			Monocytes	MC01-MC03, MC05-MC10	5060	93470	94	54	98678	0.99850	0.98944	0.99900	0.98176	0.98944	0.98559	
			Dendritic cells	DC02-DC06	65	98512	2	99	98678	0.99898	0.39634	0.99998	0.97015	0.39634	0.56277	
			NK cells	NK01-NK02, NK04-NK06	9066	89237	124	251	98678	0.99620	0.97306	0.99861	0.98651	0.97306	0.97974	
			T cells	TC01-TC14, TC16-TC21	72839	25384	318	137	98678	0.99539	0.99812	0.98763	0.99565	0.99812	0.99688	
	Step 48	added-predict	B cells	BC06	854	0	0	8	862	0.99072	0.99072	NA	1.00000	0.99072	0.99534	0.97176
	Step 49	added-predict	T cells	TC22	951	0	0	11	962	0.98857	0.98857	NA	1.00000	0.98857	0.99425	
	Step 50	added-predict	Monocytes	MC11	435	0	0	1	436	0.99771	0.99771	NA	1.00000	0.99771	0.99885	
	Step 51	added-predict	Monocytes	MC12	50	0	0	0	50	1.00000	1.00000	NA	1.00000	1.00000	1.00000	
	Step 52	added-predict	T cells	TC23	654	0	0	40	694	0.94236	0.94236	NA	1.00000	0.94236	0.97032	
	Step 53	added-predict	Dendritic cells	DC07	62	0	0	14	76	0.81579	0.81579	NA	1.00000	0.81579	0.89855	
	Step 54	added-predict	NK cells	NK07	203	0	0	16	219	0.92694	0.92694	NA	1.00000	0.92694	0.96208	
	Step 55	added-predict	pDC	DC08	26	0	0	4	30	0.86667	0.86667	NA	1.00000	0.86667	0.92857	
	Step 56	BroadS1-test	B cells	BC02	1530	11523	0	130	13183	0.99014	0.92169	1.00000	1.00000	0.92169	0.95925	0.93803
			Monocytes	MC04	1635	11430	92	26	13183	0.99105	0.98435	0.99202	0.94673	0.98435	0.96517	
			Dendritic cells	DC01	80	13025	16	62	13183	0.99408	0.56338	0.99877	0.83333	0.56338	0.67227	
			NK cells	NK03	1158	11397	392	236	13183	0.95236	0.83070	0.96675	0.74710	0.83070	0.78669	
			T cells	TC15	7963	4540	317	363	13183	0.94842	0.95640	0.93473	0.96171	0.95640	0.95905	
Cycle 7	Step 57	2-fold cross validation, (10x+GEO+BroadS2)	B cells	BC01, BC03-BC06	11949	90001	37	20	102007	0.99944	0.99833	0.99959	0.99691	0.99833	0.99762	0.99278
			Monocytes	MC01-MC03, MC05-MC12	5533	96284	123	67	102007	0.99814	0.98804	0.99872	0.97825	0.98804	0.98312	
			Dendritic cells	DC02-DC08	93	101736	1	177	102007	0.99826	0.34444	0.99999	0.98936	0.34444	0.51098	
			NK cells	NK01-NK02, NK04-NK07	9245	92308	163	291	102007	0.99555	0.96948	0.99824	0.98267	0.96948	0.97603	
			T cells	TC01-TC14, TC16-TC23	74450	26962	413	182	102007	0.99417	0.99756	0.98491	0.99448	0.99756	0.99602	
	… …	… … (future data)	… …	… …	… …	… …	… …	… …	… …	… …	… …	… …	… …	… …	… …	… …
	Step 58	BroadS1-test	B cells	BC02	1544	11520	3	116	13183	0.99097	0.93012	0.99974	0.99806	0.93012	0.96289	0.94614
			Monocytes	MC04	1615	11511	11	46	13183	0.99568	0.97231	0.99905	0.99323	0.97231	0.98266	
			Dendritic cells	DC01	136	13010	31	6	13183	0.99719	0.95775	0.99762	0.81437	0.95775	0.88026	
			NK cells	NK03	1072	11511	278	322	13183	0.95449	0.76901	0.97642	0.79407	0.76901	0.78134	
			T cells	TC15	8106	4470	387	220	13183	0.95396	0.97358	0.92032	0.95443	0.97358	0.96391	
Swapping	Step 59	2-fold cross validation, (10x+GEO+BroadS1)	B cells	BC01-BC02	11713	91103	50	32	102898	0.99920	0.99728	0.99945	0.99575	0.99728	0.99651	0.99189
			Monocytes	MC01-MC04	5041	97707	88	62	102898	0.99854	0.98785	0.99910	0.98284	0.98785	0.98534	
			Dendritic cells	DC01	71	102756	71	0	102898	0.99931	1.00000	0.99931	0.50000	1.00000	0.66667	
			NK cells	NK01-NK03	9801	92432	287	378	102898	0.99354	0.96286	0.99690	0.97155	0.96286	0.96719	
			T cells	TC01-TC15	75420	26742	374	362	102898	0.99285	0.99522	0.98621	0.99507	0.99522	0.99514	
Swapping	Step 60	BroadS2-test	B cells	BC03-BC06	1875	10269	139	9	12292	0.98796	0.99522	0.98664	0.93098	0.99522	0.96203	0.91734
			Monocytes	MC05-MC12	2123	9985	175	9	12292	0.98503	0.99578	0.98278	0.92385	0.99578	0.95847	
			Dendritic cells	DC02-DC08	0	12021	1	270	12292	0.97795	0.00000	0.99992	0.00000	0.00000	0.00000	
			NK cells	NK04-NK07	780	10826	624	62	12292	0.94419	0.92637	0.94550	0.55556	0.92637	0.69457	
			T cells	TC16-TC23	6498	5051	77	666	12292	0.93955	0.90704	0.98498	0.98829	0.90704	0.94592	
					TP	TN	FP	FN	# of Cells	ACC	SE	SP	PR	RE	F1	

❖ Supplemental Table 4. Confusion matrices for incremental learning by cycles and steps.

LEGEND

2-fold cross-validation
New set classification
Test Result (BroadS1)
Final Result - BroadS1
Final Result - BroadS2
Nan - not analyzed

CYCLES	STEPS	TRAINING SETS	TESTING SETS

Cycle 0

Step 1 | 10x dataset | 10x dataset | 2-fold cross-validation

Accuracy:	0.9987			
Precision:	0.9998	0.9837	0.9968	0.9994
Recall/Sensitivity:	0.9990	0.9922	0.9973	0.9991
Specificity:	1.0000	0.9995	0.9997	0.9981
F1_Score:	0.9994	0.9880	0.9971	0.9992

	B_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	4979	3	0	2	4984
Monocytes	0	1271	1	9	1281
NK_cells	0	2	4101	9	4112
T_cells	1	16	12	32306	32335
All-predicted	4980	1292	4114	32326	42712

Accuracy:	0.9986			
Precision:	0.9986	0.9924	0.9979	0.9989
Recall/Sensitivity:	0.9996	0.9850	0.9963	0.9993
Specificity:	0.9998	0.9998	0.9998	0.9968
F1_Score:	0.9991	0.9887	0.9971	0.9991

	B_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	5099	0	0	2	5101
Monocytes	3	1311	0	17	1331
NK_cells	0	1	4257	15	4273
T_cells	4	9	9	31984	32006
All-predicted	5106	1321	4266	32018	42711

Accuracy:	0.9987			
Precision:	0.9992	0.9881	0.9974	0.9992
Recall/Sensitivity:	0.9993	0.9886	0.9968	0.9992
Specificity:	0.9999	0.9996	0.9997	0.9974
F1_Score:	0.9993	0.9883	0.9971	0.9992

	B_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	10078	3	0	4	10085
Monocytes	3	2582	1	26	2612
NK_cells	0	3	8358	24	8385
T_cells	5	25	21	64290	64341
All-predicted	10086	2613	8380	64344	85423

Step 2 | 10x dataset | GEO_1a | New set classification

Accuracy:	0.8800		
Precision:	1.0000	0.0000	
Recall/Sensitivity:	0.8800	0.0000	
Specificity:	Nan	0.8800	
F1_Score:	0.9362	0.0000	

	Monocytes	NK_cells	All-true
Monocytes	374	51	425
All-predicted	374	51	425

Step 3 | 10x dataset | GEO_1b | New set classification

Accuracy:	0.7610		
Precision:	1.0000	0.0000	
Recall/Sensitivity:	0.7610	0.0000	
Specificity:	Nan	0.7610	
F1_Score:	0.8643	0.0000	

	Monocytes	NK_cells	All-true
Monocytes	328	103	431
All-predicted	328	103	431

Step 4 | 10x dataset | BroadS1 (test) | Test Result (BroadS1)

Accuracy:	0.8186				
Precision:	1.0000	0.0000	0.9257	0.3804	0.9956
Recall/Sensitivity:	0.8301	0.0000	0.8928	0.9878	0.7872
Specificity:	1.0000	1.0000	0.9897	0.8097	0.9940
F1_Score:	0.9072	0.0000	0.9090	0.5493	0.8792

	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	1378	0	8	262	12	1660
Dendritic_cells	0	0	111	31	0	142
Monocytes	0	0	1483	178	0	1661
NK_cells	0	0	0	1377	17	1394
T_cells	0	0	0	1772	6554	8326
All-predicted	1378	0	1602	3620	6583	13183

Cycle 1

Step 5 10x+GEO_1 10x+GEO_1 — 2-fold cross-validation

Accuracy:	0.9986			
Precision:	0.9998	0.9898	0.9974	0.9990
Recall/Sensitivity:	0.9986	0.9922	0.9957	0.9993
Specificity:	1.0000	0.9996	0.9997	0.9970
F1_Score:	0.9992	0.9910	0.9965	0.9991

	B_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	4981	1	1	5	4988
Monocytes	0	1654	1	12	1667
NK_cells	0	3	4145	15	4163
T_cells	1	13	9	32299	32322
All-predicted	4982	1671	4156	32331	43140

Accuracy:	0.9983			
Precision:	0.9988	0.9867	0.9995	0.9987
Recall/Sensitivity:	0.9992	0.9895	0.9938	0.9992
Specificity:	0.9998	0.9994	0.9999	0.9961
F1_Score:	0.9990	0.9881	0.9967	0.9989

Predicted	B_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	5093	1	0	3	5097
Monocytes	3	1782	1	15	1801
NK_cells	0	1	4196	25	4222
T_cells	3	22	1	31993	32019
All-predicted	5099	1806	4198	32036	43139

Accuracy:	0.9984			
Precision:	0.9993	0.9883	0.9984	0.9988
Recall/Sensitivity:	0.9989	0.9908	0.9948	0.9992
Specificity:	0.9999	0.9995	0.9998	0.9966
F1_Score:	0.9991	0.9895	0.9966	0.9990

Predicted	B_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	10074	2	1	8	10085
Monocytes	3	3436	2	27	3468
NK_cells	0	4	8341	40	8385
T_cells	4	35	10	64292	64341
All-predicted	10081	3477	8354	64367	86279

Step 6 10x+GEO_1 GEO_2a — New set classification

Accuracy:	1.0000	
Precision:	1.0000	
Recall/Sensitivity:	1.0000	
Specificity:	Nan	
F1_Score:	1.0000	

	NK_cells	All-true
NK_cells	309	309
All-predicted	309	309

Step 7 10x+GEO_1 GEO_2b — New set classification

Accuracy:	0.2523		
Precision:	0.0000	0.0000	1.0000
Recall/Sensitivity:	0.0000	0.0000	0.2523
Specificity:	0.3514	0.9009	Nan
F1_Score:	0.0000	0.0000	0.4029

	Monocytes	NK_cells	T_cells	All-true
T_cells	144	22	56	222
All-predicted	144	22	56	222

Step 8 10x+GEO_1 GEO_2c — New set classification

Accuracy:	0.3129		
Precision:	0.0000	0.0000	1.0000
Recall/Sensitivity:	0.0000	0.0000	0.3129
Specificity:	0.5355	0.7774	Nan
F1_Score:	0.0000	0.0000	0.4767

	Monocytes	NK_cells	T_cells	All-true
T_cells	144	69	97	310
All-predicted	144	69	97	310

Step 9 10x+GEO_1 GEO_2d — New set classification

Accuracy:	0.0185		
Precision:	0.0000	0.0000	1.0000
Recall/Sensitivity:	0.0000	0.0000	0.0185
Specificity:	0.7169	0.3015	Nan
F1_Score:	0.0000	0.0000	0.0363

	Monocytes	NK_cells	T_cells	All-true
T_cells	92	227	6	325
All-predicted	92	227	6	325

Step 10 10x+GEO_1 GEO_2e — New set classification

Accuracy:	0.0183		
Precision:	0.0000	0.0000	1.0000
Recall/Sensitivity:	0.0000	0.0000	0.0183
Specificity:	0.5733	0.4450	Nan
F1_Score:	0.0000	0.0000	0.0360

	Monocytes	NK_cells	T_cells	All-true
T_cells	163	212	7	382
All-predicted	163	212	7	382

Step 11 10x+GEO_1 GEO_2f — New set classification

Accuracy:	0.0352		
Precision:	0.0000	0.0000	1.0000
Recall/Sensitivity:	0.0000	0.0000	0.0352
Specificity:	0.8697	0.1655	Nan
F1_Score:	0.0000	0.0000	0.0680

	Monocytes	NK_cells	T_cells	All-true
T_cells	37	237	10	284
All-predicted	37	237	10	284

Step 12 10x+GEO_1 GEO_2g — New set classification

Accuracy:	0.0441		
Precision:	0.0000	0.0000	1.0000
Recall/Sensitivity:	0.0000	0.0000	0.0441
Specificity:	0.8186	0.2255	Nan
F1_Score:	0.0000	0.0000	0.0845

	Monocytes	NK_cells	T_cells	All-true
T_cells	37	158	9	204
All-predicted	37	158	9	204

Step 13 10x+GEO_1 BroadS1 (test) — Test Result (BroadS1)

Accuracy:	0.7823
Precision:	1.0000 0.0000 0.7314 0.3821 0.9930
Recall/Sensitivity:	0.6982 0.0000 1.0000 0.9835 0.7353
Specificity:	1.0000 1.0000 0.9471 0.8119 0.9911
F1_Score:	0.8223 0.0000 0.8449 0.5504 0.8449

	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	1159	0	52	422	27	1660
Dendritic_cells	0	0	142	0	0	142
Monocytes	0	0	1661	0	0	1661
NK_cells	0	0	7	1371	16	1394
T_cells	0	0	409	1795	6122	8326
All-predicted	1159	0	2271	3588	6165	13183

Cycle 2

Step 14 10x+GEO_1+2 10x+GEO_1+2 — 2-fold cross-validation

Accuracy:	0.9982
Precision:	0.9988 0.9928 0.9977 0.9984
Recall/Sensitivity:	0.9992 0.9828 0.9933 0.9994
Specificity:	0.9998 0.9997 0.9997 0.9952
F1_Score:	0.9990 0.9878 0.9955 0.9989

	B_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	4967	1	0	3	4971
Monocytes	2	1653	2	25	1682
NK_cells	1	3	4286	25	4315
T_cells	3	8	8	33171	33190
All-predicted	4973	1665	4296	33224	44158

Accuracy:	0.9980
Precision:	0.9990 0.9943 0.9961 0.9983
Recall/Sensitivity:	0.9998 0.9815 0.9929 0.9993
Specificity:	0.9999 0.9998 0.9996 0.9949
F1_Score:	0.9994 0.9879 0.9945 0.9988

	B_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	5113	0	0	1	5114
Monocytes	5	1753	3	25	1786
NK_cells	0	0	4348	31	4379
T_cells	0	10	14	32854	32878
All-predicted	5118	1763	4365	32911	44157

Accuracy:	0.9981
Precision:	0.9989 0.9936 0.9969 0.9983
Recall/Sensitivity:	0.9995 0.9821 0.9931 0.9993
Specificity:	0.9999 0.9997 0.9997 0.9951
F1_Score:	0.9992 0.9878 0.9950 0.9988

	B_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	10080	1	0	4	10085
Monocytes	7	3406	5	50	3468
NK_cells	1	3	8634	56	8694
T_cells	3	18	22	66025	66068
All-predicted	10091	3428	8661	66135	88315

Step 15 10x+GEO_1+2 GEO_3a — New set classification

Accuracy:	0.9907
Precision:	0.0000 1.0000
Recall/Sensitivity:	0.0000 0.9907
Specificity:	0.9907 Nan
F1_Score:	0.0000 0.9953

	NK_cells	T_cells	All-true
T_cells	9	956	965
All-predicted	9	956	965

Step 16 10x+GEO_1+2 GEO_3b — New set classification

Accuracy:	0.9931
Precision:	0.0000 1.0000
Recall/Sensitivity:	0.0000 0.9931
Specificity:	0.9931 Nan
F1_Score:	0.0000 0.9965

	NK_cells	T_cells	All-true
T_cells	3	432	435
All-predicted	3	432	435

Step 17 10x+GEO_1+2 BroadS1 (test) — Test Result (BroadS1)

Accuracy:	0.9222
Precision:	1.0000 0.0000 0.9098 0.8433 0.9219
Recall/Sensitivity:	0.8620 0.0000 0.9777 0.6679 0.9814
Specificity:	1.0000 1.0000 0.9860 0.9853 0.8575
F1_Score:	0.9259 0.0000 0.9425 0.7454 0.9507

	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	1431	0	22	17	190	1660
Dendritic_cells	0	0	138	0	4	142
Monocytes	0	0	1624	1	36	1661
NK_cells	0	0	1	931	462	1394
T_cells	0	0	0	155	8171	8326
All-predicted	1431	0	1785	1104	8863	13183

Cycle 3

Step 18 10x+GEO 10x+GEO 2-fold cross-validation

Accuracy:	0.9982			
Precision:	0.9990	0.9917	0.9970	0.9985
Recall/Sensitivity:	0.9994	0.9840	0.9938	0.9993
Specificity:	0.9999	0.9997	0.9997	0.9955
F1_Score:	0.9992	0.9878	0.9954	0.9989

	B_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	4981	1	0	2	4984
Monocytes	1	1663	1	25	1690
NK_cells	1	3	4302	23	4329
T_cells	3	10	12	33830	33855
All-predicted	4986	1677	4315	33880	44858

Accuracy:	0.9982			
Precision:	0.9980	0.9960	0.9961	0.9986
Recall/Sensitivity:	0.9998	0.9831	0.9943	0.9993
Specificity:	0.9997	0.9998	0.9996	0.9959
F1_Score:	0.9989	0.9895	0.9952	0.9990

	B_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	5100	0	0	1	5101
Monocytes	7	1748	3	20	1778
NK_cells	0	0	4340	25	4365
T_cells	3	7	14	33589	33613
All-predicted	5110	1755	4357	33635	44857

Accuracy:	0.9982			
Precision:	0.9985	0.9938	0.9965	0.9986
Recall/Sensitivity:	0.9996	0.9836	0.9940	0.9993
Specificity:	0.9998	0.9998	0.9996	0.9957
F1_Score:	0.9991	0.9887	0.9953	0.9989

	B_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	10081	1	0	3	10085
Monocytes	8	3411	4	45	3468
NK_cells	1	3	8642	48	8694
T_cells	6	17	26	67419	67468
All-predicted	10096	3432	8672	67515	89715

Step 19 10x+GEO BroadS2_1a New set classification

Accuracy:	0.8333			
Precision:	1.0000	0.0000	0.0000	
Recall/Sensitivity:	0.8333	0.0000	0.0000	
Specificity:	Nan	0.9722	0.8611	
F1_Score:	0.9091	0.0000	0.0000	

	B_cells	Monocytes	T_cells	All-true
B_cells	240	8	40	288
All-predicted	240	8	40	288

Step 20 10x+GEO BroadS2_1b New set classification

Accuracy:	0.9800			
Precision:	0.0000	0.0000	1.0000	
Recall/Sensitivity:	0.0000	0.0000	0.9800	
Specificity:	0.9855	0.9945	Nan	
F1_Score:	0.0000	0.0000	0.9899	

	Monocytes	NK_cells	T_cells	All-true
T_cells	8	3	539	550
All-predicted	8	3	539	550

Step 21 10x+GEO BroadS2_1c New set classification

Accuracy:	1.0000	
Precision:	1.0000	
Recall/Sensitivity:	1.0000	
Specificity:	Nan	
F1_Score:	1.0000	

	Monocytes	All-true
Monocytes	640	640
All-predicted	640	640

Step 22 10x+GEO BroadS2_1d New set classification

Accuracy:	1.0000	
Precision:	1.0000	
Recall/Sensitivity:	1.0000	
Specificity:	Nan	
F1_Score:	1.0000	

	Monocytes	All-true
Monocytes	102	102
All-predicted	102	102

Step 23 10x+GEO BroadS2_1e New set classification

Accuracy:	0.9438			
Precision:	0.0000	0.0000	1.0000	
Recall/Sensitivity:	0.0000	0.0000	0.9438	
Specificity:	0.9821	0.9617	Nan	
F1_Score:	0.0000	0.0000	0.9711	

	Monocytes	NK_cells	T_cells	All-true
T_cells	21	45	1108	1174
All-predicted	21	45	1108	1174

Step 24 10x+GEO BroadS2_1f New set classification

Accuracy:	0.0000		
Precision:	0.0000	0.0000	
Recall/Sensitivity:	0.0000	0.0000	
Specificity:	Nan	0.0000	
F1_Score:	0.0000	0.0000	

	Dendritic_cells	Monocytes	All-true
Dendritic_cells	0	55	55
All-predicted	0	55	55

Step 25 10x+GEO BroadS2_1g New set classification

Accuracy:	0.7711			
Precision:	0.0000	1.0000	0.0000	
Recall/Sensitivity:	0.0000	0.7711	0.0000	
Specificity:	0.9699	Nan	0.8012	
F1_Score:	0.0000	0.8707	0.0000	

	Monocytes	NK_cells	T_cells	All-true
NK_cells	5	128	33	166
All-predicted	5	128	33	166

Step 26 10x+GEO BroadS2_1h

New set classification

Accuracy:	0.0000		
Precision:	0.0000	0.0000	
Recall/Sensitivity:	0.0000	0.0000	
Specificity:	Nan	0.0000	
F1_Score:	0.0000	0.0000	
	Dendritic_cells	Monocytes	All-true
Dendritic_cells	0	26	26
All-predicted	0	26	26

Step 27 10x+GEO BroadS1 (test)

Test Result (BroadS1)

Accuracy:	0.9295					
Precision:	1.0000	0.0000	0.9027	0.8027	0.9428	
Recall/Sensitivity:	0.8699	0.0000	0.9946	0.7590	0.9729	
Specificity:	1.0000	1.0000	0.9846	0.9779	0.8989	
F1_Score:	0.9304	0.0000	0.9464	0.7802	0.9576	
	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	1444	0	31	37	148	1660
Dendritic_cells	0	0	142	0	0	142
Monocytes	0	0	1652	0	9	1661
NK_cells	0	0	2	1058	334	1394
T_cells	0	0	3	223	8100	8326
All-predicted	1444	0	1830	1318	8591	13183

Cycle 4

Step 28 10x+GEO+BroadS2_1 10x+GEO+BroadS2_1

2-fold cross-validation

Accuracy:	0.9964					
Precision:	0.9988	0.0000	0.9842	0.9897	0.9976	
Recall/Sensitivity:	0.9984	0.0000	0.9894	0.9876	0.9987	
Specificity:	0.9999	1.0000	0.9993	0.9989	0.9929	
F1_Score:	0.9986	0.0000	0.9868	0.9886	0.9982	
	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	5101	0	1	4	3	5109
Dendritic_cells	2	0	19	10	6	37
Monocytes	2	0	2053	0	20	2075
NK_cells	0	0	1	4309	53	4363
T_cells	2	0	12	31	34729	34774
All-predicted	5107	0	2086	4354	34811	46358

Accuracy:	0.9958					
Precision:	0.9945	0.0000	0.9822	0.9964	0.9968	
Recall/Sensitivity:	0.9998	0.0000	0.9822	0.9818	0.9992	
Specificity:	0.9993	1.0000	0.9991	0.9996	0.9907	
F1_Score:	0.9972	0.0000	0.9822	0.9890	0.9980	
	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	5263	0	0	0	1	5264
Dendritic_cells	11	0	28	1	4	44
Monocytes	10	0	2097	0	28	2135
NK_cells	3	0	1	4415	78	4497
T_cells	5	0	9	15	34389	34418
All-predicted	5292	0	2135	4431	34500	46358

Accuracy:	0.9961					
Precision:	0.9967	0.0000	0.9832	0.9930	0.9972	
Recall/Sensitivity:	0.9991	0.0000	0.9858	0.9847	0.9989	
Specificity:	0.9996	1.0000	0.9992	0.9993	0.9918	
F1_Score:	0.9979	0.0000	0.9845	0.9888	0.9981	
	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	10364	0	1	4	4	10373
Dendritic_cells	13	0	47	11	10	81
Monocytes	12	0	4150	0	48	4210
NK_cells	3	0	2	8724	131	8860
T_cells	7	0	21	46	69118	69192
All-predicted	10399	0	4221	8785	69311	92716

Step 29 10x+GEO+BroadS2_1 BroadS2_2a

New set classification

Accuracy:	0.9716		
Precision:	1.0000	0.0000	
Recall/Sensitivity:	0.9716	0.0000	
Specificity:	Nan	0.9716	
F1_Score:	0.9856	0.0000	
	B_cells	T_cells	All-true
B_cells	377	11	388
All-predicted	377	11	388

Step 30 10x+GEO+BroadS2_1 BroadS2_2b

New set classification

Accuracy:	0.9945			
Precision:	0.0000	0.0000	1.0000	
Recall/Sensitivity:	0.0000	0.0000	0.9945	
Specificity:	0.9956	0.9989	Nan	
F1_Score:	0.0000	0.0000	0.9972	
	Monocytes	NK_cells	T_cells	All-true
T_cells	4	1	903	908
All-predicted	4	1	903	908

Step 31 10x+GEO+BroadS2_1 BroadS2_2c

New set classification

Accuracy:	0.9974		
Precision:	1.0000	0.0000	
Recall/Sensitivity:	0.9974	0.0000	
Specificity:	Nan	0.9974	
F1_Score:	0.9987	0.0000	
	Monocytes	T_cells	All-true
Monocytes	378	1	379
All-predicted	378	1	379

Step 32 10x+GEO+BroadS2_1 BroadS2_2d

New set classification

Accuracy:	1.0000	
Precision:	1.0000	
Recall/Sensitivity:	1.0000	
Specificity:	Nan	
F1_Score:	1.0000	
	Monocytes	All-true
Monocytes	73	73
All-predicted	73	73

Step 33 10x+GEO+BroadS2_1 BroadS2_2e

New set classification

Accuracy:	0.9874		
Precision:	0.0000	0.0000	1.0000
Recall/Sensitivity:	0.0000	0.0000	0.9874
Specificity:	0.9979	0.9895	Nan
F1_Score:	0.0000	0.0000	0.9937

	Monocytes	NK_cells	T_cells	All-true
T_cells	2	10	942	954
All-predicted	2	10	942	954

Step 34 10x+GEO+BroadS2_1 BroadS2_2f

New set classification

Accuracy:	0.7273		
Precision:	1.0000	0.0000	
Recall/Sensitivity:	0.7273	0.0000	
Specificity:	Nan	0.7273	
F1_Score:	0.8421	0.0000	

	Dendritic_cells	Monocytes	All-true
Dendritic_cells	24	9	33
All-predicted	24	9	33

Step 35 10x+GEO+BroadS2_1 BroadS2_2g

New set classification

Accuracy:	0.4297			
Precision:	0.0000	1.0000	0.0000	
Recall/Sensitivity:	0.0000	0.4297	0.0000	
Specificity:	0.9962	Nan	0.4335	
F1_Score:	0.0000	0.6011	0.0000	

	Dendritic_cells	NK_cells	T_cells	All-true
NK_cells	1	113	149	263
All-predicted	1	113	149	263

Step 36 10x+GEO+BroadS2_1 BroadS2_2h

New set classification

Accuracy:	0.9167		
Precision:	1.0000	0.0000	
Recall/Sensitivity:	0.9167	0.0000	
Specificity:	Nan	0.9167	
F1_Score:	0.9565	0.0000	

	Dendritic_cells	T_cells	All-true
Dendritic_cells	11	1	12
All-predicted	11	1	12

Step 37 10x+GEO+BroadS2_1 BroadS1 (test)

Test Result (BroadS1)

Accuracy:	0.9312					
Precision:	0.9987	0.8257	0.9249	0.8678	0.9293	
Recall/Sensitivity:	0.9042	0.6338	0.9856	0.6119	0.9843	
Specificity:	0.9998	0.9985	0.9885	0.9890	0.8717	
F1_Score:	0.9491	0.7171	0.9542	0.7177	0.9560	

	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	1501	15	82	0	62	1660
Dendritic_cells	0	90	49	0	3	142
Monocytes	0	4	1637	0	20	1661
NK_cells	2	0	1	853	538	1394
T_cells	0	0	1	130	8195	8326
All-predicted	1503	109	1770	983	8818	13183

Cycle 5

Step 38 10x+GEO+BroadS2_1+2 10x+GEO+BroadS2_1+2

2-fold cross-validation

Accuracy:	0.9956					
Precision:	0.9992	0.9744	0.9852	0.9932	0.9960	
Recall/Sensitivity:	0.9981	0.6230	0.9882	0.9751	0.9989	
Specificity:	0.9999	1.0000	0.9993	0.9993	0.9883	
F1_Score:	0.9987	0.7600	0.9867	0.9841	0.9974	

	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	5318	0	1	1	8	5328
Dendritic_cells	0	38	22	0	1	61
Monocytes	2	0	2257	1	24	2284
NK_cells	1	1	0	4392	110	4504
T_cells	1	0	11	28	35646	35686
All-predicted	5322	39	2291	4422	35789	47863

Accuracy:	0.9952					
Precision:	0.9983	0.9697	0.9800	0.9914	0.9963	
Recall/Sensitivity:	0.9987	0.4923	0.9882	0.9777	0.9984	
Specificity:	0.9998	1.0000	0.9989	0.9991	0.9895	
F1_Score:	0.9985	0.6531	0.9841	0.9845	0.9973	

	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	5426	0	1	0	6	5433
Dendritic_cells	2	32	29	0	2	65
Monocytes	3	1	2350	0	24	2378
NK_cells	1	0	3	4516	99	4619
T_cells	3	0	15	39	35311	35368
All-predicted	5435	33	2398	4555	35442	47863

Accuracy:	0.9954					
Precision:	0.9988	0.9720	0.9826	0.9923	0.9962	
Recall/Sensitivity:	0.9984	0.5576	0.9882	0.9764	0.9986	
Specificity:	0.9998	1.0000	0.9991	0.9992	0.9889	
F1_Score:	0.9986	0.7065	0.9854	0.9843	0.9974	

	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	10744	0	2	1	14	10761
Dendritic_cells	2	70	51	0	3	126
Monocytes	5	1	4607	1	48	4662
NK_cells	2	1	3	8908	209	9123
T_cells	4	0	26	67	70957	71054
All-predicted	10757	72	4689	8977	71231	95726

Step 39 10x+GEO+BroadS2_1+2 BroadS2_3a

New set classification

Accuracy:	0.9942		
Precision:	1.0000	0.0000	
Recall/Sensitivity:	0.9942	0.0000	
Specificity:	Nan	0.9942	
F1_Score:	0.9971	0.0000	
	B_cells	T_cells	All-true
B_cells	344	2	346
All-predicted	344	2	346

Step 40 10x+GEO+BroadS2_1+2 BroadS2_3b

New set classification

Accuracy:	0.9854		
Precision:	0.0000	1.0000	
Recall/Sensitivity:	0.0000	0.9854	
Specificity:	0.9854	Nan	
F1_Score:	0.0000	0.9927	
	NK_cells	T_cells	All-true
T_cells	14	946	960
All-predicted	14	946	960

Step 41 10x+GEO+BroadS2_1+2 BroadS2_3c

New set classification

Accuracy:	0.9972		
Precision:	1.0000	0.0000	
Recall/Sensitivity:	0.9972	0.0000	
Specificity:	Nan	0.9972	
F1_Score:	0.9986	0.0000	
	Monocytes	T_cells	All-true
Monocytes	353	1	354
All-predicted	353	1	354

Step 42 10x+GEO+BroadS2_1+2 BroadS2_3d

New set classification

Accuracy:	1.0000	
Precision:	1.0000	
Recall/Sensitivity:	1.0000	
Specificity:	Nan	
F1_Score:	1.0000	
	Monocytes	All-true
Monocytes	98	98
All-predicted	98	98

Step 43 10x+GEO+BroadS2_1+2 BroadS2_3e

New set classification

Accuracy:	0.9751		
Precision:	0.0000	1.0000	
Recall/Sensitivity:	0.0000	0.9751	
Specificity:	0.9751	Nan	
F1_Score:	0.0000	0.9874	
	NK_cells	T_cells	All-true
T_cells	24	938	962
All-predicted	24	938	962

Step 44 10x+GEO+BroadS2_1+2 BroadS2_3f

New set classification

Accuracy:	0.7895				
Precision:	0.0000	1.0000	0.0000	0.0000	
Recall/Sensitivity:	0.0000	0.7895	0.0000	0.0000	
Specificity:	0.9737	Nan	0.9737	0.8421	
F1_Score:	0.0000	0.8824	0.0000	0.0000	
	B_cells	Dendritic_cells	Monocytes	T_cells	All-true
Dendritic_cells	1	30	1	6	38
All-predicted	1	30	1	6	38

Step 45 10x+GEO+BroadS2_1+2 BroadS2_3g

New set classification

Accuracy:	0.7835		
Precision:	1.0000	0.0000	
Recall/Sensitivity:	0.7835	0.0000	
Specificity:	Nan	0.7835	
F1_Score:	0.8786	0.0000	
	NK_cells	T_cells	All-true
NK_cells	152	42	194
All-predicted	152	42	194

Step 46 10x+GEO+BroadS2_1+2 BroadS1 (test)

Test Result (BroadS1)

Accuracy:	0.9354					
Precision:	0.9961	0.4000	0.9612	0.7542	0.9510	
Recall/Sensitivity:	0.9193	0.0423	0.9693	0.7769	0.9737	
Specificity:	0.9995	0.9993	0.9944	0.9701	0.9139	
F1_Score:	0.9561	0.0764	0.9652	0.7654	0.9622	
	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	1526	9	6	80	39	1660
Dendritic_cells	0	6	58	49	29	142
Monocytes	4	0	1610	5	42	1661
NK_cells	2	0	1	1083	308	1394
T_cells	0	0	0	219	8107	8326
All-predicted	1532	15	1675	1436	8525	13183

Cycle 6

Step 47 10x+GEO+BroadS2_1+2+3 10x+GEO+BroadS2_1+2+3

2-fold cross-validation

Accuracy:	0.9954					
Precision:	0.9991	0.9623	0.9882	0.9889	0.9962	
Recall/Sensitivity:	0.9985	0.7846	0.9894	0.9754	0.9983	
Specificity:	0.9999	1.0000	0.9994	0.9989	0.9890	
F1_Score:	0.9988	0.8644	0.9888	0.9821	0.9972	
	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	5492	0	1	0	7	5500
Dendritic_cells	0	51	12	0	2	65
Monocytes	4	1	2510	2	20	2537
NK_cells	0	1	2	4472	110	4585
T_cells	1	0	15	48	36588	36652
All-predicted	5497	53	2540	4522	36727	49339

Accuracy:	0.9933				
Precision:	0.9973	1.0000	0.9755	0.9841	0.9951
Recall/Sensitivity:	0.9984	0.1414	0.9895	0.9708	0.9980
Specificity:	0.9997	1.0000	0.9986	0.9983	0.9862
F1_Score:	0.9979	0.2478	0.9825	0.9774	0.9965

	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	5598	0	1	0	8	5607
Dendritic_cells	5	14	52	13	15	99
Monocytes	5	0	2550	1	21	2577
NK_cells	2	0	1	4594	135	4732
T_cells	3	0	10	60	36251	36324
All-predicted	5613	14	2614	4668	36430	49339

Accuracy:	0.9943				
Precision:	0.9982	0.9811	0.9819	0.9865	0.9957
Recall/Sensitivity:	0.9985	0.4630	0.9894	0.9731	0.9981
Specificity:	0.9998	1.0000	0.9990	0.9986	0.9876
F1_Score:	0.9983	0.5561	0.9856	0.9798	0.9969

	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	11090	0	2	0	15	11107
Dendritic_cells	5	65	64	13	17	164
Monocytes	9	1	5060	3	41	5114
NK_cells	2	1	3	9066	245	9317
T_cells	4	0	25	108	72839	72976
All-predicted	11110	67	5154	9190	73157	98678

Step 48 10x+GEO+BroadS2_1+2+3 BroadS2_4a

New set classification

Accuracy:	0.9907				
Precision:	1.0000	0.0000	0.0000	0.0000	
Recall/Sensitivity:	0.9907	0.0000	0.0000	0.0000	
Specificity:	Nan	0.9988	0.9954	0.9965	
F1_Score:	0.9953	0.0000	0.0000	0.0000	

	B_cells	Dendritic_cells	Monocytes	T_cells	All-true
B_cells	854	1	4	3	862
All-predicted	854	1	4	3	862

Step 49 10x+GEO+BroadS2_1+2+3 BroadS2_4b

New set classification

Accuracy:	0.9886					
Precision:	0.0000	0.0000	0.0000	0.0000	1.0000	
Recall/Sensitivity:	0.0000	0.0000	0.0000	0.0000	0.9886	
Specificity:	0.9979	0.9969	0.9958	0.9979	Nan	
F1_Score:	0.0000	0.0000	0.0000	0.0000	0.9942	

	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
T_cells	2	3	4	2	951	962
All-predicted	2	3	4	2	951	962

Step 50 10x+GEO+BroadS2_1+2+3 BroadS2_4c

New set classification

Accuracy:	0.9977		
Precision:	0.0000	1.0000	
Recall/Sensitivity:	0.0000	0.9977	
Specificity:	0.9977	Nan	
F1_Score:	0.0000	0.9989	

	Dendritic_cells	Monocytes	All-true
Monocytes	1	435	436
All-predicted	1	435	436

Step 51 10x+GEO+BroadS2_1+2+3 BroadS2_4d

New set classification

Accuracy:	1.0000
Precision:	1.0000
Recall/Sensitivity:	1.0000
Specificity:	Nan
F1_Score:	1.0000

	Monocytes	All-true
Monocytes	50	50
All-predicted	50	50

Step 52 10x+GEO+BroadS2_1+2+3 BroadS2_4e

New set classification

Accuracy:	0.9424			
Precision:	0.0000	0.0000	1.0000	
Recall/Sensitivity:	0.0000	0.0000	0.9424	
Specificity:	0.9986	0.9438	Nan	
F1_Score:	0.0000	0.0000	0.9703	

	B_cells	NK_cells	T_cells	All-true
T_cells	1	39	654	694
All-predicted	1	39	654	694

Step 53 10x+GEO+BroadS2_1+2+3 BroadS2_4f

New set classification

Accuracy:	0.8158			
Precision:	1.0000	0.0000	0.0000	
Recall/Sensitivity:	0.8158	0.0000	0.0000	
Specificity:	Nan	0.8289	0.9868	
F1_Score:	0.8986	0.0000	0.0000	

	Dendritic_cells	Monocytes	T_cells	All-true
Dendritic_cells	62	13	1	76
All-predicted	62	13	1	76

Step 54 10x+GEO+BroadS2_1+2+3 BroadS2_4g

New set classification

Accuracy:	0.9269			
Precision:	0.0000	1.0000	0.0000	
Recall/Sensitivity:	0.0000	0.9269	0.0000	
Specificity:	0.9954	Nan	0.9315	
F1_Score:	0.0000	0.9621	0.0000	

	Monocytes	NK_cells	T_cells	All-true
NK_cells	1	203	15	219
All-predicted	1	203	15	219

Step 55 10x+GEO+BroadS2_1+2+3 BroadS2_4h

New set classification

Accuracy:	0.8667			
Precision:	1.0000	0.0000	0.0000	
Recall/Sensitivity:	0.8667	0.0000	0.0000	
Specificity:	Nan	0.9000	0.9667	
F1_Score:	0.9286	0.0000	0.0000	

	Dendritic_cells	Monocytes	T_cells	All-true
Dendritic_cells	26	3	1	30
All-predicted	26	3	1	30

Step 56 10x+GEO+BroadS2_1+2+3 BroadS1 (test)

Test Result (BroadS1)

Accuracy:	0.9380					
Precision:	1.0000	0.8333	0.9467	0.7471	0.9617	
Recall/Sensitivity:	0.9217	0.5634	0.9843	0.8307	0.9564	
Specificity:	1.0000	0.9988	0.9920	0.9667	0.9347	
F1_Score:	0.9592	0.6723	0.9652	0.7867	0.9591	
	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	1530	13	26	30	61	1660
Dendritic_cells	0	80	61	0	1	142
Monocytes	0	3	1635	0	23	1661
NK_cells	0	0	4	1158	232	1394
T_cells	0	0	1	362	7963	8326
All-predicted	1530	96	1727	1550	8280	13183

Cycle 7

Step 57 10x+GEO+BroadS2 10x+GEO+BroadS2

2-fold cross-validation

Accuracy:	0.9936					
Precision:	0.9970	1.0000	0.9772	0.9832	0.9956	
Recall/Sensitivity:	0.9981	0.4426	0.9921	0.9732	0.9974	
Specificity:	0.9996	1.0000	0.9987	0.9983	0.9878	
F1_Score:	0.9975	0.6136	0.9846	0.9782	0.9965	
	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	5895	0	3	0	8	5906
Dendritic_cells	10	54	37	7	14	122
Monocytes	3	0	2749	0	19	2771
NK_cells	1	0	1	4571	124	4697
T_cells	4	0	23	71	37410	37508
All-predicted	5913	54	2813	4649	37575	51004

Accuracy:	0.9919					
Precision:	0.9969	0.9750	0.9792	0.9821	0.9933	
Recall/Sensitivity:	0.9985	0.2635	0.9841	0.9659	0.9977	
Specificity:	0.9996	1.0000	0.9988	0.9982	0.9821	
F1_Score:	0.9977	0.4149	0.9817	0.9740	0.9955	
	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	6054	0	0	0	9	6063
Dendritic_cells	6	39	46	16	41	148
Monocytes	7	0	2784	1	37	2829
NK_cells	2	0	2	4674	161	4839
T_cells	4	1	11	68	37040	37124
All-predicted	6073	40	2843	4759	37288	51003

Accuracy:	0.9928					
Precision:	0.9969	0.9875	0.9782	0.9827	0.9945	
Recall/Sensitivity:	0.9983	0.3531	0.9881	0.9695	0.9976	
Specificity:	0.9996	1.0000	0.9987	0.9982	0.9850	
F1_Score:	0.9976	0.5143	0.9831	0.9761	0.9960	
	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	11949	0	3	0	17	11969
Dendritic_cells	16	93	83	23	55	270
Monocytes	10	0	5533	1	56	5600
NK_cells	3	0	3	9245	285	9536
T_cells	8	1	34	139	74450	74632
All-predicted	11986	94	5656	9408	74863	102007

Step 58 10x+GEO+BroadS2 BroadS1 (test)

Final Result - BroadS1

Accuracy:	0.9461					
Precision:	0.9981	0.8144	0.9932	0.7941	0.9544	
Recall/Sensitivity:	0.9301	0.9577	0.9723	0.7690	0.9736	
Specificity:	0.9997	0.9976	0.9990	0.9764	0.9203	
F1_Score:	0.9629	0.8803	0.9827	0.7813	0.9639	
	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	1544	20	5	60	31	1660
Dendritic_cells	0	136	5	0	1	142
Monocytes	1	9	1615	0	36	1661
NK_cells	2	0	1	1072	319	1394
T_cells	0	2	0	218	8106	8326
All-predicted	1547	167	1626	1350	8493	13183

Swapping

Step 59 10x+GEO+BroadS1 10x+GEO+BroadS1

2-fold cross-validation

Accuracy:	0.9918					
Precision:	0.9967	1.0000	0.9885	0.9643	0.9949	
Recall/Sensitivity:	0.9957	0.6133	0.9842	0.9693	0.9954	
Specificity:	0.9996	1.0000	0.9994	0.9961	0.9855	
F1_Score:	0.9962	0.7603	0.9864	0.9668	0.9952	
	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	5764	0	4	10	11	5789
Dendritic_cells	1	46	18	5	5	75
Monocytes	11	0	2498	0	29	2538
NK_cells	2	0	2	4831	149	4984
T_cells	5	0	5	164	37889	38063
All-predicted	5783	46	2527	5010	38083	51449

Accuracy:	0.9916					
Precision:	0.9948	1.0000	0.9872	0.9615	0.9955	
Recall/Sensitivity:	0.9988	0.3731	0.9815	0.9737	0.9947	
Specificity:	0.9993	1.0000	0.9993	0.9957	0.9878	
F1_Score:	0.9968	0.5435	0.9843	0.9676	0.9951	
	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	5949	0	0	4	3	5956
Dendritic_cells	10	25	19	12	1	67
Monocytes	13	0	2543	2	33	2591
NK_cells	2	0	1	4970	131	5104
T_cells	6	0	13	181	37531	37731
All-predicted	5980	25	2576	5169	37699	51449

Accuracy:	0.9917					
Precision:	0.9958	1.0000	0.9879	0.9629	0.9952	
Recall/Sensitivity:	0.9973	0.4932	0.9829	0.9715	0.9951	
Specificity:	0.9995	1.0000	0.9994	0.9959	0.9866	
F1_Score:	0.9965	0.6519	0.9854	0.9672	0.9951	
	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	11713	0	4	14	14	11745
Dendritic_cells	11	71	37	17	6	142
Monocytes	24	0	5041	2	62	5129
NK_cells	4	0	3	9801	280	10088
T_cells	11	0	18	345	75420	75794
All-predicted	11763	71	5103	10179	75782	102898

Step 60 10x+GEO+BroadS1 BroadS2 (test)

Final Result - BroadS2

Accuracy:	0.9173					
Precision:	0.9310	0.0000	0.9238	0.5556	0.9883	
Recall/Sensitivity:	0.9952	0.0000	0.9958	0.9264	0.9070	
Specificity:	0.9866	0.9999	0.9828	0.9455	0.9850	
F1_Score:	0.9620	0.0000	0.9585	0.6946	0.9459	
	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All-true
B_cells	1875	0	6	0	3	1884
Dendritic_cells	103	0	152	0	15	270
Monocytes	6	0	2123	0	3	2132
NK_cells	6	0	0	780	56	842
T_cells	24	1	17	624	6498	7164
All-predicted	2014	1	2298	1404	6575	12292

❖ Supplemental Table 5. The assessment of classification performance for specific simulations EXP 1 through EXP 8.

EXP	4 Supersets	Training Set	Testing Set
1 (Cycle 7)	**10x**	✓	
	GEO	✓	
	BroadS1		✓
	BroadS2	✓	

Accuracy:	0.9461					
Precision:	0.9981	0.8144	0.9932	0.7941	0.9544	
Recall/Sensitivity:	0.9301	0.9577	0.9723	0.7690	0.9736	
Specificity:	0.9997	0.9976	0.9990	0.9764	0.9203	
F1_Score:	0.9629	0.8803	0.9827	0.7813	0.9639	
Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All (true)
B_cells	1544	20	5	60	31	1660
Dendritic_cells	0	136	5	0	1	142
Monocytes	1	9	1615	0	36	1661
NK_cells	2	0	1	1072	319	1394
T_cells	0	2	0	218	8106	8326
All (predicted)	1547	167	1626	1350	8493	13183

EXP	4 Supersets	Training Set	Testing Set
2a (swapping)	**10x**	✓	
	GEO	✓	
	BroadS1	✓	
	BroadS2		✓

Accuracy:	0.9173					
Precision:	0.9310	0.0000	0.9238	0.5556	0.9883	
Recall/Sensitivity:	0.9952	0.0000	0.9958	0.9264	0.9070	
Specificity:	0.9866	0.9999	0.9828	0.9455	0.9850	
F1_Score:	0.9620	0.0000	0.9585	0.6946	0.9459	
Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All (true)
B_cells	1875	0	6	0	3	1884
Dendritic_cells	103	0	152	0	15	270
Monocytes	6	0	2123	0	3	2132
NK_cells	6	0	0	780	56	842
T_cells	24	1	17	624	6498	7164
All (predicted)	2014	1	2298	1404	6575	12292

EXP	4 Supersets	Training Set	Testing Set
2b (swapping) with QC	**10x**	✓	
	GEO	✓	
	BroadS1	✓	
	BroadS2 (QC)		✓

Accuracy:	0.9172					
Precision:	0.9317	0.0000	0.9216	0.5560	0.9884	
Recall/Sensitivity:	0.9957	0.0000	0.9965	0.9264	0.9079	
Specificity:	0.9867	1.0000	0.9832	0.9449	0.9848	
F1_Score:	0.9627	0.0000	0.9576	0.6949	0.9464	
Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All (true)
B_cells	1869	0	5	0	3	1877
Dendritic_cells	103	0	152	0	15	270
Monocytes	5	0	1997	0	2	2004
NK_cells	6	0	0	780	56	842
T_cells	23	0	13	623	6493	7152
All (predicted)	2006	0	2167	1403	6569	12145

EXP	4 Supersets	Training Set	Testing Set
3	**10x**		✓
	GEO	✓	
	BroadS1	✓	
	BroadS2	✓	

Accuracy:	0.9829					
Precision:	0.9769	0.0000	0.8493	0.9851	0.9921	
Recall/Sensitivity:	0.9616	0.0000	0.8978	0.9255	0.9972	
Specificity:	0.9970	0.9978	0.9950	0.9985	0.9759	
F1_Score:	0.9692	0.0000	0.8729	0.9544	0.9947	
Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All (true)
B_cells	9698	28	356	1	2	10085
Dendritic_cells	NA	NA	NA	NA	NA	NA
Monocytes	202	19	2345	3	43	2612
NK_cells	0	135	27	7760	463	8385
T_cells	27	7	33	113	64161	64341
All (predicted)	9927	189	2761	7877	64669	85423

EXP	4 Supersets	Training Set	Testing Set
4	10x	✓	
	GEO		✓
	BroadS1	✓	
	BroadS2	✓	

Accuracy:	0.9352					
Precision:	0.0000	0.0000	0.9976	0.5394	0.9965	
Recall/Sensitivity:	0.0000	0.0000	0.9801	0.9968	0.9169	
Specificity:	0.9993	1.0000	0.9994	0.9340	0.9914	
F1_Score:	0.0000	0.0000	0.9888	0.7000	0.9550	
Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All (true)
B_cells	NA	NA	NA	NA	NA	NA
Dendritic_cells	NA	NA	NA	NA	NA	NA
Monocytes	3	0	839	5	9	856
NK_cells	0	0	0	308	1	309
T_cells	0	0	2	258	2867	3127
All (predicted)	3	0	841	571	2877	4292

EXP	2 Sets	Training Set	Testing Set
5	BroadS1		✓
	BroadS2	✓	

Accuracy:	0.9447					
Precision:	1.0000	0.8609	0.9837	0.8150	0.9488	
Recall/Sensitivity:	0.9102	0.9155	0.9825	0.7712	0.9736	
Specificity:	1.0000	0.9984	0.9977	0.9793	0.9100	
F1_Score:	0.9530	0.8874	0.9831	0.7925	0.9611	
Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All (true)
B_cells	1511	14	15	25	95	1660
Dendritic_cells	0	130	9	0	3	142
Monocytes	0	7	1632	0	22	1661
NK_cells	0	0	2	1075	317	1394
T_cells	0	0	1	219	8106	8326
All (predicted)	1511	151	1659	1319	8543	13183

EXP	2 Sets	Training Set	Testing Set
6	BroadS1	✓	
	BroadS2		✓

Accuracy:	0.8815					
Precision:	0.9230	0.0000	0.9482	0.4363	0.9681	
Recall/Sensitivity:	0.8339	0.0000	0.9953	0.9192	0.8889	
Specificity:	0.9874	1.0000	0.9886	0.9127	0.9590	
F1_Score:	0.8762	0.0000	0.9712	0.5917	0.9268	
Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All (true)
B_cells	1571	0	7	204	102	1884
Dendritic_cells	125	0	100	15	30	270
Monocytes	0	0	2122	0	10	2132
NK_cells	0	0	0	774	68	842
T_cells	6	0	9	781	6368	7164
All (predicted)	1702	0	2238	1774	6578	12292

EXP	4 Sets	Training Set	Testing Set
7	10x	✓	
	GEO	✓	
	BroadS1		
	BroadS2		✓

Accuracy:	0.9232					
Precision:	1.0000	NA	0.8315	0.8102	0.9502	
Recall/Sensitivity:	0.8769	NA	1.0000	0.7553	0.9671	
Specificity:	1.0000	NA	0.9575	0.9870	0.9292	
F1_Score:	0.9344	NA	0.9080	0.7818	0.9586	
Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All (true)
B_cells	1652	NA	57	1	174	1884
Dendritic_cells	0	NA	265	0	5	270
Monocytes	0	NA	2132	0	0	2132
NK_cells	0	NA	22	636	184	842
T_cells	0	NA	88	148	6928	7164
All (predicted)	1652	NA	2564	785	7291	12292

EXP	4 Sets	Training Set	Testing Set
8 (Cycle 3)	10x	✓	
	GEO	✓	
	BroadS1		✓
	BroadS2		

Accuracy:	0.9295					
Precision:	1.0000	NA	0.9027	0.8027	0.9428	
Recall/Sensitivity:	0.8699	NA	0.9946	0.7590	0.9729	
Specificity:	1.0000	NA	0.9846	0.9779	0.8989	
F1_Score:	0.9304	NA	0.9464	0.7802	0.9576	
	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All (true)
B_cells	1444	NA	31	37	148	1660
Dendritic_cells	0	NA	142	0	0	142
Monocytes	0	NA	1652	0	9	1661
NK_cells	0	NA	2	1058	334	1394
T_cells	0	NA	3	223	8100	8326
All (predicted)	1444	NA	1830	1318	8591	13183

❖ Other Supplemental Materials in Study III.

- Raw data table of overall accuracy in incremental learning cycles.

OVERALL ACCURACY	Cross Validation	Added Data	External Validation	Total cells	Added cells
Cycle 0	0.99865	0.82009	0.81863	85423	0
Cycle 1	0.99842	0.24263	0.78230	86279	856
Cycle 2	0.99808	0.99143	0.92217	88315	2036
Cycle 3	0.99819	0.91869	0.92953	89715	1400
Cycle 4	0.99612	0.93721	0.93120	92716	3001
Cycle 5	0.99540	0.96917	0.93545	95726	3010
Cycle 6	0.99435	0.972	0.93803	98678	2952
Cycle 7	0.993	0	0.946	102007	3329
Swapping	0.992	0	0.917	102898	0

- Raw data tables of other assessment metrics values for each cell type of testing steps in cycles.

B cell	ACC	F1	SE	SP	PR	RE
Step 4	0.97861	0.9072	0.8301	1.0000	1.0000	0.8301
Step 13	0.96200	0.8223	0.6982	1.0000	1.0000	0.6982
Step 17	0.98263	0.92591	0.8621	1.0000	1.0000	0.8621
Step 27	0.98362	0.93041	0.8699	1.0000	1.0000	0.8699
Step 37	0.98779	0.9491	0.9042	0.9998	0.9987	0.9042
Step 46	0.98938	0.9561	0.9193	0.9995	0.9961	0.9193
Step 56	0.99014	0.9593	0.9217	1.0000	1.0000	0.9217
Step 58	0.99097	0.9629	0.9301	0.9997	0.9981	0.9301
Step 60*	0.98796	0.9620	0.9952	0.9866	0.9310	0.9952

DC	ACC	F1	SE	SP	PR	RE
Step 4	0.00000	0.0000	0.0000	NA	NA	0.0000
Step 13	0.00000	0.0000	0.0000	NA	NA	0.0000
Step 17	0.00000	0.0000	0.0000	NA	NA	0.0000
Step 27	0.0000	0.0000	0.0000	NA	NA	0.0000
Step 37	0.99461	0.7171	0.6338	0.9985	0.8257	0.6338
Step 46	0.98900	0.0764	0.0423	0.9993	0.4000	0.0423
Step 56	0.99408	0.6723	0.5634	0.9988	0.8333	0.5634
Step 58	0.99719	0.8803	0.9578	0.9976	0.8144	0.9578
Step 60*	0.97795	0.0000	0.0000	0.9999	0.0000	0.0000

Monocyte	ACC	F1	SE	SP	PR	RE
Step 4	0.97747	0.9090	0.8928	0.9897	0.9257	0.8928
Step 13	0.95373	0.8449	1.0000	0.9471	0.7314	1.0000
Step 17	0.9850	0.9425	0.9777	0.9860	0.9098	0.9777
Step 27	0.98582	0.9464	0.9946	0.9846	0.9027	0.9946
Step 37	0.98809	0.9542	0.9856	0.9885	0.9249	0.9856
Step 46	0.99120	0.9652	0.9693	0.9944	0.9612	0.9693
Step 56	0.99105	0.9652	0.9844	0.9920	0.9467	0.9844
Step 58	0.99568	0.9827	0.9723	0.9991	0.9932	0.9723
Step 60*	0.98503	0.9585	0.9958	0.9828	0.9239	0.9958

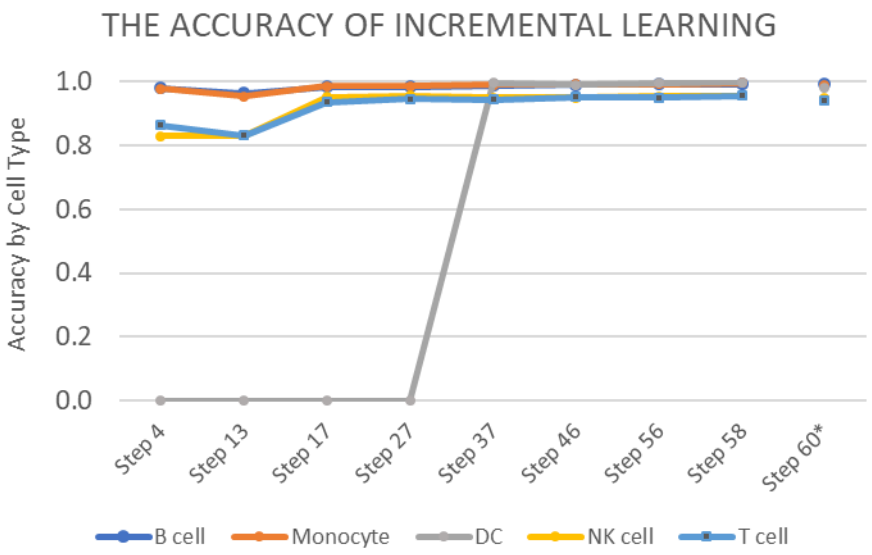
NK cell	ACC	F1	SE	SP	PR	RE
Step 4	0.82857	0.5493	0.9878	0.8097	0.3804	0.9878
Step 13	0.83008	0.5504	0.9835	0.8119	0.3821	0.9835
Step 17	0.95176	0.7454	0.6679	0.9853	0.8433	0.6679
Step 27	0.95479	0.7802	0.7590	0.9780	0.8027	0.7590
Step 37	0.94910	0.7177	0.6119	0.9890	0.8678	0.6119
Step 46	0.94963	0.7654	0.7769	0.9701	0.7542	0.7769
Step 56	0.95236	0.7867	0.8307	0.9668	0.7471	0.8307
Step 58	0.95449	0.7813	0.7690	0.9764	0.7941	0.7690
Step 60*	0.94419	0.6946	0.9264	0.9455	0.5556	0.9264

T cell	ACC	F1	SE	SP	PR	RE
Step 4	0.86338	0.8792	0.7872	0.9940	0.9956	0.7872
Step 13	0.82955	0.8449	0.7353	0.9912	0.9930	0.7353
Step 17	0.93575	0.9507	0.9814	0.8575	0.9219	0.9814
Step 27	0.94561	0.9576	0.9729	0.8989	0.9429	0.9729
Step 37	0.94281	0.9560	0.9843	0.8717	0.9294	0.9843
Step 46	0.95168	0.9622	0.9737	0.9139	0.9510	0.9737
Step 56	0.94842	0.9591	0.9564	0.9347	0.9617	0.9564
Step 58	0.95396	0.9639	0.9736	0.9203	0.9544	0.9736
Step 60*	0.93955	0.9459	0.9070	0.9850	0.9883	0.9070

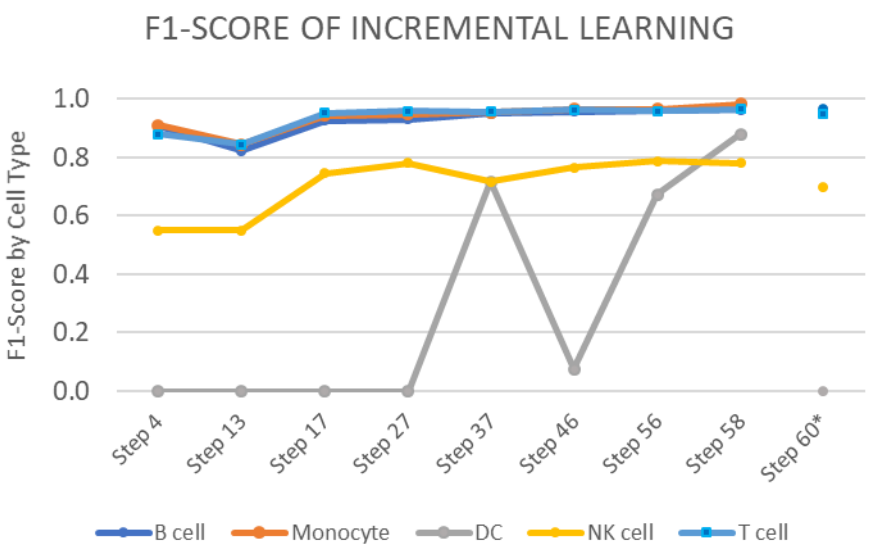
ACC	B cell	Monocyte	DC	NK cell	T cell
Step 4	0.97861	0.97747	0.00000	0.82857	0.86338
Step 13	0.96200	0.95373	0.00000	0.83008	0.82955
Step 17	0.98263	0.9850	0.00000	0.95176	0.93575
Step 27	0.98362	0.98582	0.0000	0.95479	0.94561
Step 37	0.98779	0.98809	0.99461	0.94910	0.94281
Step 46	0.98938	0.99120	0.98900	0.94963	0.95168
Step 56	0.99014	0.99105	0.99408	0.95236	0.94842
Step 58	0.99097	0.99568	0.99719	0.95449	0.95396
Step 60*	0.98796	0.98503	0.97795	0.94419	0.93955

F1	B cell	Monocyte	DC	NK cell	T cell
Step 4	0.9072	0.9090	0.0000	0.5493	0.8792
Step 13	0.8223	0.8449	0.0000	0.5504	0.8449
Step 17	0.92591	0.9425	0.0000	0.7454	0.9507
Step 27	0.93041	0.9464	0.0000	0.7802	0.9576
Step 37	0.9491	0.9542	0.7171	0.7177	0.9560
Step 46	0.9561	0.9652	0.0764	0.7654	0.9622
Step 56	0.9593	0.9652	0.6723	0.7867	0.9591
Step 58	0.9629	0.9827	0.8803	0.7813	0.9639
Step 60*	0.9620	0.9585	0.0000	0.6946	0.9459

- The accuracy of each cell type of testing steps during incremental learning cycles.



- The F1 score of each cell type of testing steps during incremental learning cycles.



- Raw data tables of confusion matrix values for each cell type of testing steps in incremental learning cycles.

		TP	TN	FP	FN	Total#
B cells	Step 4	1378	11523	0	282	13183
	Step 13	1159	11523	0	501	13183
	Step 17	1431	11523	0	229	13183
	Step 27	1444	11523	0	216	13183
	Step 37	1501	11521	2	159	13183
	Step 46	1526	11517	6	134	13183
	Step 56	1530	11523	0	130	13183
	Step 58	1544	11520	3	116	13183
	swapping	1875	10269	139	9	12292

		TP	TN	FP	FN	Total#
Monocytes	Step 4	1483	11403	119	178	13183
	Step 13	1661	10912	610	0	13183
	Step 17	1624	11361	161	37	13183
	Step 27	1652	11344	178	9	13183
	Step 37	1637	11389	133	24	13183
	Step 46	1610	11457	65	51	13183
	Step 56	1635	11430	92	26	13183
	Step 58	1615	11511	11	46	13183
	swapping	2123	9985	175	9	12292

		TP	TN	FP	FN	Total#
Dendritic cells	Step 4	0	0	0	142	
	Step 13	0	0	0	142	
	Step 17	0	0	0	142	
	Step 27	0	0	0	142	
	Step 37	90	13022	19	52	13183
	Step 46	6	13032	9	136	13183
	Step 56	80	13025	16	62	13183
	Step 58	136	13010	31	6	13183
	swapping	0	12021	1	270	12292

		TP	TN	FP	FN	Total#
NK cells	Step 4	1377	9546	2243	17	13183
	Step 13	1371	9572	2217	23	13183
	Step 17	931	11616	173	463	13183
	Step 27	1058	11529	260	336	13183
	Step 37	853	11659	130	541	13183
	Step 46	1083	11436	353	311	13183
	Step 56	1158	11397	392	236	13183
	Step 58	1072	11511	278	322	13183
	swapping	780	10826	624	62	12292

		TP	TN	FP	FN	Total#
T cells	Step 4	6554	4828	29	1772	13183
	Step 13	6122	4814	43	2204	13183
	Step 17	8171	4165	692	155	13183
	Step 27	8100	4366	491	226	13183
	Step 37	8195	4234	623	131	13183
	Step 46	8107	4439	418	219	13183
	Step 56	7963	4540	317	363	13183
	Step 58	8106	4470	387	220	13183
	swapping	6498	5051	77	666	12292

- Raw data tables of confusion matrix values in each cell type of cross validation and added prediction in cycles.

B cells		TP	TN	FP	FN	Total#
(2-fold)	Step 1	10078	75330	8	7	85423
(2-fold)	Step 5	10074	76187	7	11	86279
(2-fold)	Step 14	10080	78219	11	5	88315
(2-fold)	Step 18	10081	79615	15	4	89715
(added-predict-BC)	Step 19	240	0	0	48	288
(2-fold)	Step 28	10364	82308	35	9	92716
(added-predict-BC)	Step 29	377	0	0	11	388
(2-fold)	Step 38	10744	84952	13	17	95726
(added-predict-BC)	Step 39	344	0	0	2	346
(2-fold)	Step 47	11090	87551	20	17	98678
(added-predict-BC)	Step 48	854	0	0	8	862
(2-fold)	Step 57	11949	90001	37	20	102007
Swapping	Step 59	11713	91103	50	32	102898

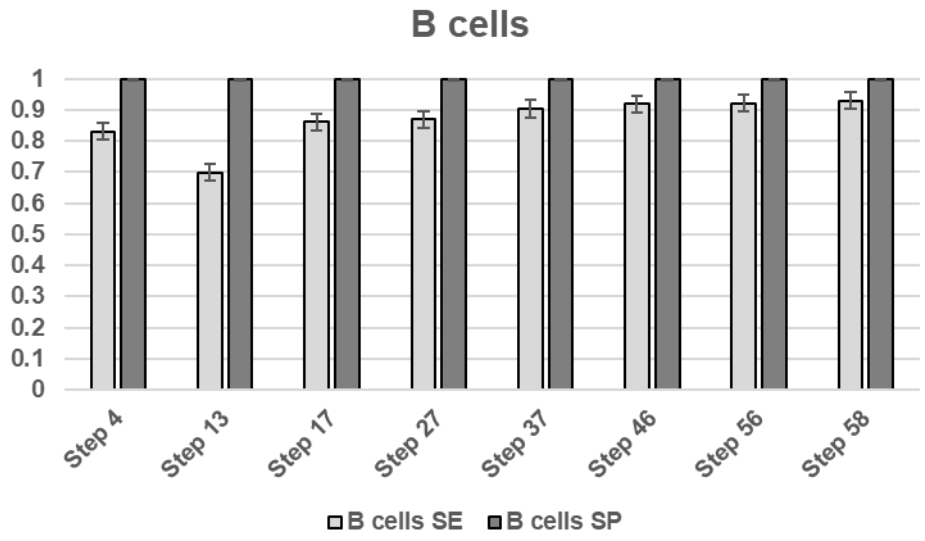
Monocytes		TP	TN	FP	FN	Total#
(2-fold)	Step 1	2582	82780	31	30	85423
(added-predict-MC)	Step 2	374	0	0	51	425
(added-predict-MC)	Step 3	328	0	0	103	431
(2-fold)	Step 5	3436	82770	41	32	86279
(2-fold)	Step 14	3406	84825	22	62	88315
(2-fold)	Step 18	3411	86226	21	57	89715
(added-predict-MC)	Step 21	640	0	0	0	640
(added-predict-MC)	Step 22	102	0	0	0	102
(2-fold)	Step 28	4150	88435	71	60	92716
(added-predict-MC)	Step 31	378	0	0	1	379
(added-predict-MC)	Step 32	73	0	0	0	73
(2-fold)	Step 38	4607	90982	82	55	95726
(added-predict-MC)	Step 41	353	0	0	1	354
(added-predict-MC)	Step 42	98	0	0	0	98
(2-fold)	Step 47	5060	93470	94	54	98678
(added-predict-MC)	Step 50	435	0	0	1	436
(added-predict-MC)	Step 51	50	0	0	0	50
(2-fold)	Step 57	5533	96284	123	67	102007
Swapping	Step 59	5041	97707	88	62	102898

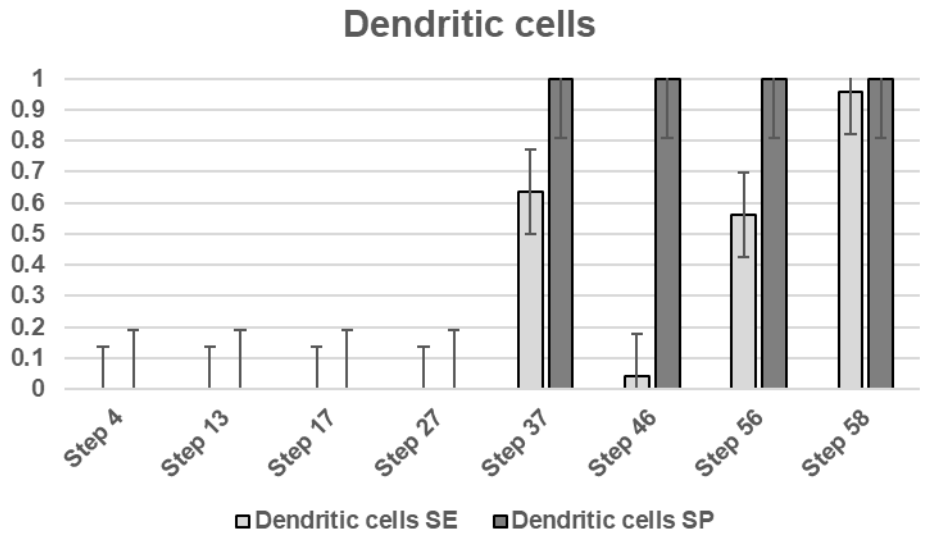
Dendritic cells		TP	TN	FP	FN	Total#
(added-predict-DC)	Step 24	0	0	0	55	55
(added-predict-DC)	Step 26	0	0	0	26	26
(2-fold)	Step 28	0	92635	0	81	92716
(added-predict-DC)	Step 34	24	0	0	9	33
(added-predict-DC)	Step 36	11	0	0	1	12
(2-fold)	Step 38	70	95598	2	56	95726
(added-predict-DC)	Step 44	30	0	0	8	38
(2-fold)	Step 47	65	98512	2	99	98678
(added-predict-DC)	Step 53	62	0	0	14	76
(added-predict-DC)	Step 55	26	0	0	4	30
(2-fold)	Step 57	93	101736	1	177	102007
Swapping	Step 59	71	102756	71	0	102898

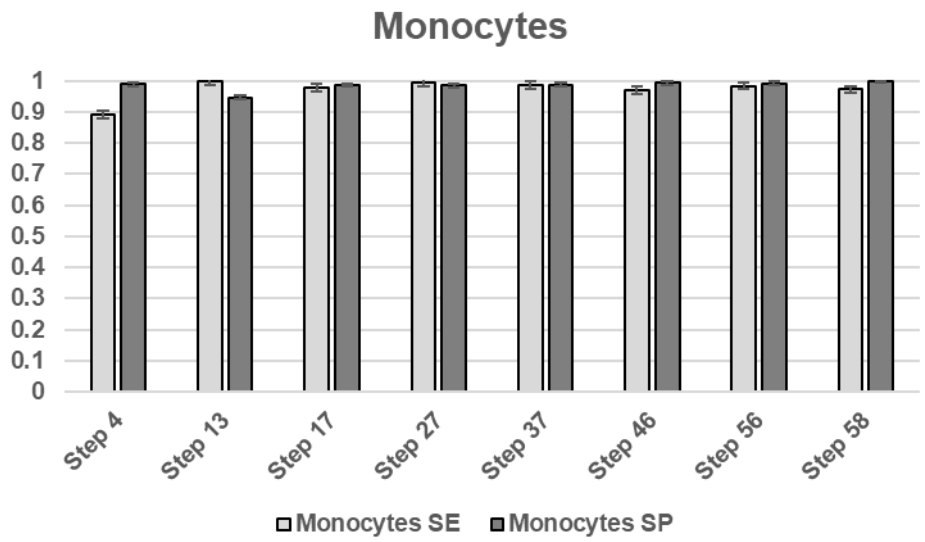
NK cells		TP	TN	FP	FN	Total#
(2-fold)	Step 1	8358	77016	22	27	85423
(2-fold)	Step 5	8341	77881	13	44	86279
(added-predict-NK)	Step 6	309	0	0	0	309
(2-fold)	Step 14	8634	79594	27	60	88315
(2-fold)	Step 18	8642	80991	30	52	89715
(added-predict-NK)	Step 25	128	0	0	38	166
(2-fold)	Step 28	8724	83795	61	136	92716
(added-predict-NK)	Step 35	113	0	0	150	263
(2-fold)	Step 38	8908	86534	69	215	95726
(added-predict-NK)	Step 45	152	0	0	42	194
(2-fold)	Step 47	9066	89237	124	251	98678
(added-predict-NK)	Step 54	203	0	0	16	219
(2-fold)	Step 57	9245	92308	163	291	102007
Swapping	Step 59	9801	92432	287	378	102898

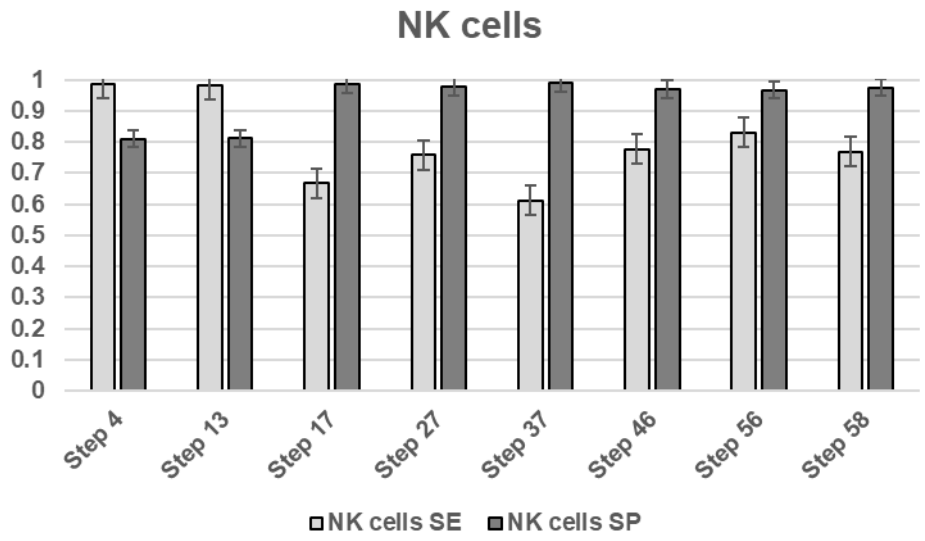
T cells		TP	TN	FP	FN	Total#
(2-fold)	Step 1	64290	21028	54	51	85423
(2-fold)	Step 5	64292	21863	75	49	86279
(added-predict-TC)	Step 7	56	0	0	166	222
(added-predict-TC)	Step 8	97	0	0	213	310
(added-predict-TC)	Step 9	6	0	0	319	325
(added-predict-TC)	Step 10	7	0	0	375	382
(added-predict-TC)	Step 11	10	0	0	274	284
(added-predict-TC)	Step 12	9	0	0	195	204
(2-fold)	Step 14	66025	22137	110	43	88315
(added-predict-TC)	Step 15	956	0	0	9	965
(added-predict-TC)	Step 16	432	0	0	3	435
(2-fold)	Step 18	67419	22151	96	49	89715
(added-predict-TC)	Step 20	539	0	0	11	550
(added-predict-TC)	Step 23	1108	0	0	66	1174
(2-fold)	Step 28	69118	23331	193	74	92716
(added-predict-TC)	Step 30	903	0	0	5	908
(added-predict-TC)	Step 33	942	0	0	12	954
(2-fold)	Step 38	70957	24398	274	97	95726
(added-predict-TC)	Step 40	946	0	0	14	960
(added-predict-TC)	Step 43	938	0	0	24	962
(2-fold)	Step 47	72839	25384	318	137	98678
(added-predict-TC)	Step 49	951	0	0	11	962
(added-predict-TC)	Step 52	654	0	0	40	694
(2-fold)	Step 57	74450	26962	413	182	102007
Swapping	Step 59	75420	26742	374	362	102898

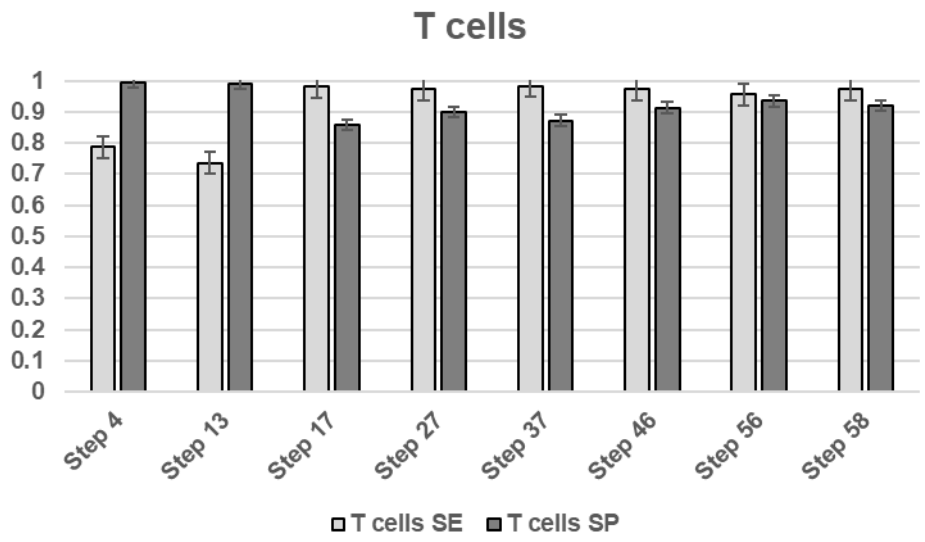
- ANN predication performance (SE and SP) on each cell type (B cells, Monocytes, NK cells, T cells, and Dendritic cells) in the incremental learning experiment.






Monocytes
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Step 4
Step 13
Step 17
Step 27
Step 37
Step 46
Step 56
Step 58
Monocytes SE
Monocytes SP


NK cells
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Step 4
Step 13
Step 17
Step 27
Step 37
Step 46
Step 56
Step 58
NK cells SE
NK cells SP


T cells
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Step 4
Step 13
Step 17
Step 27
Step 37
Step 46
Step 56
Step 58
T cells SE
T cells SP

Appendix 8 Raw Results in Study IV

- Raw results of confusion matrix during 17 rounds of four-supersets-swapping external cross-validation experiments.

Round1-AllSets+10*5EC

TestWith-Source-BroadS

Accuracy:	0.933323				
Precision:	0.99934	0.570776	0.97976	0.753272	0.955041
Recall/Ser	0.911446	0.880282	0.932571	0.825681	0.956762
Specificity	0.999913	0.992792	0.997223	0.968021	0.922792
F1_Score:	0.953371	0.692521	0.955583	0.787817	0.955901

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1513	25	16	18	88	1660
Dendritic	0	125	14	0	3	142
Monocyte	0	68	1549	0	44	1661
NK_cells	1	0	2	1151	240	1394
T_cells	0	1	0	359	7966	8326
All	1514	219	1581	1528	8341	13183

TestWith-Source-BroadS

Accuracy:	0.897169				
Precision:	0.976719	0.592593	0.92548	0.492395	0.963328
Recall/Ser	0.957537	0.059259	0.972796	0.922803	0.887353
Specificity	0.995869	0.999085	0.983563	0.930044	0.952808
F1_Score:	0.967033	0.107744	0.948548	0.642149	0.923781

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1804	0	32	0	48	1884
Dendritic	15	16	116	11	112	270
Monocyte	17	10	2074	11	20	2132
NK_cells	0	0	3	777	62	842
T_cells	11	1	16	779	6357	7164
All	1847	27	2241	1578	6599	12292

TestWith-Source-10x

Accuracy:	0.059281			
Precision:	0.686747	0.03145	1	0.96851
Recall/Ser	0.005652	0.997703	0.000239	0.037286
Specificity	0.999655	0.030865	1	0.9963
F1_Score:	0.011212	0.060978	0.000477	0.071807

Predicted	B_cells	Monocytes	NK_cells	T_cells	All
B_cells	57	10028	0	0	10085
Monocyte	1	2606	0	5	2612
NK_cells	20	8290	2	73	8385
T_cells	5	61937	0	2399	64341
All	83	82861	2	2477	85423

TestWith-Source-GEODB

Accuracy:	0.751758				
Precision:	0.674718	0.196078	0.36102	0.099451	0.961145
Recall/Ser	0.699889	0.002293	0.731199	0.965517	0.886721
Specificity	0.981583	0.998649	0.863487	0.91888	0.908766
F1_Score:	0.687073	0.004532	0.483378	0.180328	0.922434

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	11	204	79	245	1796
Dendritic	132	10	3748	176	296	4362
Monocyte	64	10	2421	465	351	3311
NK_cells	0	10	0	308	1	319
T_cells	410	10	333	2069	22090	24912
All	1863	51	6706	3097	22983	34700

Round2-AllSets+5*5EC

Accuracy:	0.940605				
Precision:	0.996154	0.80597	0.95399	0.776259	0.956932
Recall/Ser	0.936145	0.760563	0.986153	0.774032	0.963368
Specificity	0.999479	0.998006	0.993144	0.973619	0.925674
F1_Score:	0.965217	0.782609	0.969805	0.775144	0.960139

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1554	17	36	16	37	1660
Dendritic	0	108	32	0	2	142
Monocyte	4	8	1638	0	11	1661
NK_cells	2	0	2	1079	311	1394
T_cells	0	1	9	295	8021	8326
All	1560	134	1717	1390	8382	13183

Accuracy:	0.934429				
Precision:	0.963141	0.780612	0.943675	0.73057	0.950041
Recall/Ser	0.957006	0.566667	0.958724	0.669834	0.96622
Specificity	0.99337	0.996423	0.987992	0.981834	0.929017
F1_Score:	0.960064	0.656652	0.95114	0.698885	0.958062

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1803	1	22	0	58	1884
Dendritic	22	153	86	1	8	270
Monocyte	24	39	2044	0	25	2132
NK_cells	4	0	1	564	273	842
T_cells	19	3	13	207	6922	7164
All	1872	196	2166	772	7286	12292

Accuracy:	0.14509				
Precision:	0.300725	0	0.028666	0.997672	0.928715
Recall/Ser	0.01646	0	0.812021	0.102206	0.143765
Specificity	0.994876	0.999262	0.132144	0.999974	0.966322
F1_Score:	0.031212	0	0.055378	0.185418	0.248987

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	166	2	9914	0	3	10085
Monocyte	2	59	2121	0	430	2612
NK_cells	13	1	7237	857	277	8385
T_cells	371	1	54717	2	9250	64341
All	552	63	73989	859	9960	85423

Accuracy:	0.752156				
Precision:	0.674718	0.192308	0.36102	0.099451	0.961145
Recall/Ser	0.701843	0.001148	0.732305	0.980892	0.886899
Specificity	0.981572	0.999307	0.8634	0.918832	0.908579
F1_Score:	0.688013	0.002282	0.48362	0.180592	0.922531

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	6	204	79	245	1791
Dendritic	132	5	3748	176	296	4357
Monocyte	64	5	2421	465	351	3306
NK_cells	0	5	0	308	1	314
T_cells	410	5	333	2069	22090	24907
All	1863	26	6706	3097	22983	34675

Round3-AllSets+2*5EC

Accuracy:	0.931882				
Precision:	0.99605	0.698324	0.985267	0.749307	0.946267
Recall/Ser	0.911446	0.880282	0.966285	0.776184	0.956041
Specificity	0.999479	0.995859	0.997917	0.969293	0.906938
F1_Score:	0.951872	0.778816	0.975684	0.762509	0.951129

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1513	17	7	0	123	1660
Dendritic_	0	125	14	0	3	142
Monocyte	1	37	1605	0	18	1661
NK_cells	2	0	2	1082	308	1394
T_cells	3	0	1	362	7960	8326
All	1519	179	1629	1444	8412	13183

Accuracy:	0.891555				
Precision:	0.936056	0.860465	0.830196	0.558603	0.961133
Recall/Ser	0.924628	0.137037	0.951689	0.7981	0.904383
Specificity	0.988566	0.999501	0.959154	0.953624	0.948908
F1_Score:	0.930307	0.236422	0.886801	0.657213	0.931895

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1742	1	121	4	16	1884
Dendritic_	19	37	80	1	133	270
Monocyte	62	2	2029	0	39	2132
NK_cells	16	0	80	672	74	842
T_cells	22	3	134	526	6479	7164
All	1861	43	2444	1203	6741	12292

Accuracy:	0.053896				
Precision:	0.433526	0	0.031056	1	0.56944
Recall/Ser	0.02231	0	0.970904	0.000239	0.028613
Specificity	0.996098	0.999895	0.044523	1	0.933972
F1_Score:	0.042437	0	0.060186	0.000477	0.054488

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	225	0	9856	0	4	10085
Monocyte	7	6	2536	0	63	2612
NK_cells	33	2	7023	2	1325	8385
T_cells	254	1	62245	0	1841	64341
All	519	9	81660	2	3233	85423

Accuracy:	0.752395				
Precision:	0.674718	0.181818	0.36102	0.099451	0.961145
Recall/Ser	0.70302	0.000459	0.73297	0.990354	0.887006
Specificity	0.981565	0.999703	0.863348	0.918804	0.908467
F1_Score:	0.688578	0.000916	0.483765	0.180751	0.922589

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	3	204	79	245	1788
Dendritic_	132	2	3748	176	296	4354
Monocyte	64	2	2421	465	351	3303
NK_cells	0	2	0	308	1	311
T_cells	410	2	333	2069	22090	24904
All	1863	11	6706	3097	22983	34660

Round4-AllSets+1*5EC

Accuracy:	0.936964				
Precision:	0.991525	0.661202	0.980296	0.821718	0.94145
Recall/Ser	0.916265	0.852113	0.958459	0.727403	0.973337
Specificity	0.998872	0.995246	0.997223	0.981339	0.896232
F1_Score:	0.952411	0.744615	0.969254	0.77169	0.957128

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1521	37	6	0	96	1660
Dendritic_	0	121	19	0	2	142
Monocyte	11	25	1592	0	33	1661
NK_cells	2	0	5	1014	373	1394
T_cells	0	0	2	220	8104	8326
All	1534	183	1624	1234	8608	13183

Accuracy:	0.898226				
Precision:	0.976164	0.90625	0.942492	0.463306	0.964746
Recall/Ser	0.934713	0.214815	0.968574	0.862233	0.897683
Specificity	0.995869	0.999501	0.987598	0.92655	0.954173
F1_Score:	0.954989	0.347305	0.955355	0.60274	0.930007

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1761	0	5	18	100	1884
Dendritic_	23	58	90	90	9	270
Monocyte	5	6	2065	35	21	2132
NK_cells	0	0	11	726	105	842
T_cells	15	0	20	698	6431	7164
All	1804	64	2191	1567	6666	12292

Accuracy:	0.081828				
Precision:	0.404506	0	0.031378	1	0.794106
Recall/Ser	0.037382	0	0.952527	0.000239	0.06408
Specificity	0.992633	0.999941	0.072539	1	0.949293
F1_Score:	0.06844	0	0.060754	0.000477	0.118591

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	377	0	9702	0	6	10085
Monocyte	8	5	2488	0	111	2612
NK_cells	3	0	7428	2	952	8385
T_cells	544	0	59674	0	4123	64341
All	932	5	79292	2	5192	85423

Accuracy:	0.752474				
Precision:	0.674718	0.166667	0.36102	0.099451	0.961145
Recall/Ser	0.703414	0.00023	0.733192	0.993548	0.887042
Specificity	0.981563	0.999835	0.86333	0.918795	0.908429
F1_Score:	0.688767	0.000459	0.483813	0.180804	0.922608

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	2	204	79	245	1787
Dendritic_	132	1	3748	176	296	4353
Monocyte	64	1	2421	465	351	3302
NK_cells	0	1	0	308	1	310
T_cells	410	1	333	2069	22090	24903
All	1863	6	6706	3097	22983	34655

Round5-r'1*5EC

Accuracy:	0.936888				
Precision:	0.996078	0.801527	0.947093	0.796467	0.947765
Recall/Ser	0.918072	0.739437	0.980735	0.743902	0.967571
Specificity	0.999479	0.998006	0.992102	0.977521	0.908586
F1_Score:	0.955486	0.769231	0.96362	0.769288	0.957566

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1524	16	42	2	76	1660
Dendritic_	0	105	35	0	2	142
Monocyte	4	9	1629	0	19	1661
NK_cells	2	0	8	1037	347	1394
T_cells	0	1	6	263	8056	8326
All	1530	131	1720	1302	8500	13183

Accuracy:	0.910023				
Precision:	0.94925	0.714286	0.898032	0.671218	0.938041
Recall/Ser	0.873673	0.240741	0.941839	0.758907	0.953099
Specificity	0.991545	0.997837	0.977559	0.972664	0.912051
F1_Score:	0.909895	0.360111	0.919414	0.712375	0.94551

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1646	10	45	0	183	1884
Dendritic_	17	65	166	4	18	270
Monocyte	62	12	2008	0	50	2132
NK_cells	0	1	2	639	200	842
T_cells	9	3	15	309	6828	7164
All	1734	91	2236	952	7279	12292

Accuracy:	0.128162				
Precision:	0.499219	0	0.031808	1	0.945082
Recall/Ser	0.095092	0	0.918836	0.030769	0.11394
Specificity	0.987231	0.99959	0.117847	1	0.979793
F1_Score:	0.159753	0	0.061488	0.059701	0.203362

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	959	0	9121	0	5	10085
Monocyte	3	34	2400	0	175	2612
NK_cells	6	0	7875	258	246	8385
T_cells	953	1	56056	0	7331	64341
All	1921	35	75452	258	7757	85423

Accuracy:	0.752554				
Precision:	0.674718	0	0.36102	0.099451	0.961145
Recall/Ser	0.703807	0	0.733414	0.996764	0.887077
Specificity	0.98156	0.999967	0.863313	0.918785	0.908391
F1_Score:	0.688956	0	0.483861	0.180857	0.922627

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	1	204	79	245	1786
Dendritic_	132	0	3748	176	296	4352
Monocyte	64	0	2421	465	351	3301
NK_cells	0	0	0	308	1	309
T_cells	410	0	333	2069	22090	24902
All	1863	1	6706	3097	22983	34650

Round6-r'tumor_DC

Accuracy:	0.911553				
Precision:	0.945075	0	0.936738	0.675258	0.937819
Recall/Ser	0.953614	0	0.971704	0.657819	0.949195
Specificity	0.992016	1	0.99054	0.962592	0.892114
F1_Score:	0.949325	0	0.953901	0.666424	0.943473

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1583	0	25	26	26	1660
Dendritic_	65	0	76	0	1	142
Monocyte	14	0	1614	0	33	1661
NK_cells	10	0	3	917	464	1394
T_cells	3	0	5	415	7903	8326
All	1675	0	1723	1358	8427	13183

Accuracy:	0.932232				
Precision:	0.946623	0.823077	0.940514	0.797637	0.940377
Recall/Ser	0.922505	0.396296	0.978893	0.64133	0.975293
Specificity	0.990584	0.998087	0.987008	0.988035	0.913612
F1_Score:	0.934409	0.535	0.95932	0.710994	0.957517

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1738	9	31	1	105	1884
Dendritic_	70	107	78	1	14	270
Monocyte	9	5	2087	0	31	2132
NK_cells	1	2	6	540	293	842
T_cells	18	7	17	135	6987	7164
All	1836	130	2219	677	7430	12292

Accuracy:	0.072077				
Precision:	0.325893	0	0.028899	1	0.903218
Recall/Ser	1.45E-02	0	0.892802	0.000239	5.71E-02
Specificity	0.995991	0.997588	0.053701	1	0.981311
F1_Score:	0.027722	0	0.055985	0.000477	0.107496

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	146	1	9936	0	2	10085
Monocyte	9	140	2332	0	131	2612
NK_cells	0	35	8087	2	261	8385
T_cells	293	30	60341	0	3677	64341
All	448	206	80696	2	4071	85423

Accuracy:	0.789297				
Precision:	0.711778	0	0.451006	0.104442	0.962401
Recall/Ser	0.703807	0	0.733414	0.996764	0.887077
Specificity	0.983713	0.999967	0.900895	0.919305	0.893915
F1_Score:	0.70777	0	0.558542	0.189073	0.923206

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	1	204	79	245	1786
Dendritic_	35	0	2410	28	266	2739
Monocyte	64	0	2421	465	351	3301
NK_cells	0	0	0	308	1	309
T_cells	410	0	333	2069	22090	24902
All	1766	1	5368	2949	22953	33037

Round7-r'tonsil_DC

Accuracy:	0.941136				
Precision:	0.99737	0.83871	0.970238	0.743742	0.962932
Recall/Ser	0.913855	0.915493	0.981337	0.8099	0.960966
Specificity	0.999653	0.998083	0.99566	0.967003	0.936586
F1_Score:	0.953788	0.875421	0.975756	0.775412	0.961948

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1517	14	37	67	25	1660
Dendritic_	0	130	10	0	2	142
Monocyte	1	10	1630	0	20	1661
NK_cells	2	0	2	1129	261	1394
T_cells	1	1	1	322	8001	8326
All	1521	155	1680	1518	8309	13183

Accuracy:	0.929873				
Precision:	0.983778	0.96	0.947628	0.699264	0.941354
Recall/Ser	0.901274	0.355556	0.992964	0.789786	0.956728
Specificity	0.99731	0.999667	0.988484	0.975022	0.916732
F1_Score:	0.94072	0.518919	0.969766	0.741774	0.948979

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1698	0	10	2	174	1884
Dendritic_	19	96	89	1	65	270
Monocyte	0	2	2117	0	13	2132
NK_cells	2	0	0	665	175	842
T_cells	7	2	18	283	6854	7164
All	1726	100	2234	951	7281	12292

Accuracy:	0.198401				
Precision:	0.465875	0.028737	1	0.914184	
Recall/Ser	1.56E-02	0.75804	0.020751	0.227491	
Specificity	0.997611	0.191883	1	0.934826	
F1_Score:	0.030129	0.055375	0.040659	0.364322	

Predicted	B_cells	Monocytes	NK_cells	T_cells	All
B_cells	157	9906	0	22	10085
Monocyte	6	1980	0	626	2612
NK_cells	1	7484	174	726	8385
T_cells	173	49531	0	14637	64341
All	337	68901	174	16011	85423

Accuracy:	0.860651				
Precision:	0.72617	0	0.818458	0.105443	0.973685
Recall/Ser	0.703807	0	0.733414	0.996764	0.887077
Specificity	0.983375	0.999967	0.980109	0.912868	0.889362
F1_Score:	0.714814	0	0.773606	0.190712	0.928366

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	1	204	79	245	1786
Monocyte	64	0	2421	465	351	3301
NK_cells	0	0	0	308	1	309
T_cells	410	0	333	2069	22090	24902
All	1731	1	2958	2921	22687	30298

Round8-r'methanol_T8

Accuracy:	0.936509				
Precision:	0.996053	0.611111	0.973292	0.777385	0.953811
Recall/Ser	0.912048	0.929577	0.943408	0.789096	0.964809
Specificity	0.999479	0.993559	0.996268	0.97328	0.919909
F1_Score:	0.952201	0.73743	0.958117	0.783197	0.959279

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1514	16	30	30	70	1660
Dendritic_	0	132	7	0	3	142
Monocyte	1	67	1567	0	26	1661
NK_cells	2	0	2	1100	290	1394
T_cells	3	1	4	285	8033	8326
All	1520	216	1610	1415	8422	13183

Accuracy:	0.913358				
Precision:	0.964365	0.886792	0.870632	0.619782	0.961799
Recall/Ser	0.919321	0.174074	0.981707	0.811164	0.931323
Specificity	0.993851	0.999501	0.96939	0.963406	0.948323
F1_Score:	0.941304	0.291022	0.92284	0.702675	0.946316

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1732	0	51	0	101	1884
Dendritic_	8	47	204	0	11	270
Monocyte	26	3	2093	0	10	2132
NK_cells	3	0	13	683	143	842
T_cells	27	3	43	419	6672	7164
All	1796	53	2404	1102	6937	12292

Accuracy:	0.127589				
Precision:	0.838028	0	0.031333	0.833333	0.949289
Recall/Ser	3.54E-02	0	0.916539	0.000596	1.27E-01
Specificity	0.999084	0.999895	0.106278	0.999987	0.979366
F1_Score:	0.067929	0	0.060595	0.001192	0.223344

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	357	1	9720	0	7	10085
Monocyte	25	8	2394	1	184	2612
NK_cells	24	0	8112	5	244	8385
T_cells	20	0	56178	0	8143	64341
All	426	9	76404	6	8578	85423

Accuracy:	0.890194				
Precision:	0.812016	0	0.846504	0.172549	0.969149
Recall/Ser	0.703807	0	0.733414	0.996764	0.930766
Specificity	0.987752	0.999961	0.980264	0.941473	0.889362
F1_Score:	0.754049	0	0.785911	0.294174	0.94957

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	1	204	79	245	1786
Monocyte	64	0	2421	465	351	3301
NK_cells	0	0	0	308	1	309
T_cells	227	0	235	933	18754	20149
All	1548	1	2860	1785	19351	25545

Round9-r'IL_10_T4_d1

Accuracy:	0.938785				
Precision:	0.996667	0.702128	0.963702	0.772161	0.95749
Recall/Ser	0.900602	0.929577	0.959061	0.799857	0.96577
Specificity	0.999566	0.995706	0.994793	0.972093	0.926498
F1_Score:	0.946203	0.8	0.961376	0.785765	0.961612

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1495	14	51	46	54	1660
Dendritic_	0	132	7	0	3	142
Monocyte	1	40	1593	0	27	1661
NK_cells	3	1	2	1115	273	1394
T_cells	1	1	0	283	8041	8326
All	1500	188	1653	1444	8398	13183

Accuracy:	0.907419				
Precision:	0.972425	0.869565	0.862696	0.584769	0.95783
Recall/Ser	0.917197	0.148148	0.978424	0.766033	0.92895
Specificity	0.995292	0.999501	0.967323	0.96	0.942863
F1_Score:	0.944004	0.253165	0.916923	0.663239	0.943169

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1728	4	108	17	27	1884
Dendritic_	17	40	150	3	60	270
Monocyte	23	0	2086	2	21	2132
NK_cells	0	0	12	645	185	842
T_cells	9	2	62	436	6655	7164
All	1777	46	2418	1103	6948	12292

Accuracy:	0.217389				
Precision:	0.749655	0	0.034255	0.991892	0.95405
Recall/Ser	1.08E-01	0	0.893185	0.043769	2.30E-01
Specificity	0.995182	0.999988	0.205734	0.999961	0.966227
F1_Score:	0.18847	0	0.065979	0.083838	0.370334

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1087	1	8982	0	15	10085
Monocyte	15	0	2333	2	262	2612
NK_cells	0	0	7583	367	435	8385
T_cells	348	0	49209	1	14783	64341
All	1450	1	68107	370	15495	85423

Accuracy:	0.884805				
Precision:	0.812016	0	0.846504	0.173131	0.967035
Recall/Ser	0.703807	0	0.733414	0.996764	0.926516
Specificity	0.987074	0.999959	0.979092	0.93868	0.889362
F1_Score:	0.754049	0	0.785911	0.295019	0.946342

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	1	204	79	245	1786
Monocyte	64	0	2421	465	351	3301
NK_cells	0	0	0	308	1	309
T_cells	227	0	235	927	17513	18902
All	1548	1	2860	1779	18110	24298

Round10-r'IL-10_T4_d2

Accuracy:	0.942047				
Precision:	0.992935	0.685864	0.9895	0.792398	0.953598
Recall/Ser	0.931325	0.922535	0.964479	0.777618	0.967571
Specificity	0.999045	0.995399	0.998525	0.97591	0.919292
F1_Score:	0.961144	0.786787	0.976829	0.784938	0.960534

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1546	43	5	17	49	1660
Dendritic_	0	131	9	0	2	142
Monocyte	9	16	1602	0	34	1661
NK_cells	2	0	1	1084	307	1394
T_cells	0	1	2	267	8056	8326
All	1557	191	1619	1368	8448	13183

Accuracy:	0.866173				
Precision:	0.989785	0	0.911803	0.372631	0.989087
Recall/Ser	0.977176	0	0.989212	0.957245	0.822306
Specificity	0.998174	1	0.979921	0.881485	0.987324
F1_Score:	0.98344	0	0.948931	0.536439	0.898018

Predicted	B_cells	Monocytes	NK_cells	T_cells	All
B_cells	1841	3	11	29	1884
Dendritic_	4	159	103	4	270
Monocyte	3	2109	14	6	2132
NK_cells	1	9	806	26	842
T_cells	11	33	1229	5891	7164
All	1860	2313	2163	5956	12292

Accuracy:	0.095677				
Precision:	0.828423	0	0.029671	1	0.916967
Recall/Ser	4.74E-02	0	0.896248	0.001073	8.31E-02
Specificity	0.998686	0.998724	0.075509	1	0.977042
F1_Score:	0.089664	0	0.05744	0.002144	0.152344

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	478	0	9599	0	8	10085
Monocyte	9	96	2341	0	166	2612
NK_cells	12	3	8051	9	310	8385
T_cells	78	10	58908	0	5345	64341
All	577	109	78899	9	5829	85423

Accuracy:	0.875603				
Precision:	0.81254	0	0.846504	0.174307	0.963196
Recall/Ser	0.703807	0	0.733414	0.996764	0.919059
Specificity	0.985929	0.999955	0.97701	0.933943	0.889362
F1_Score:	0.754275	0	0.785911	0.296724	0.94061

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	1	204	79	245	1786
Monocyte	64	0	2421	465	351	3301
NK_cells	0	0	0	308	1	309
T_cells	226	0	235	915	15624	17000
All	1547	1	2860	1767	16221	22396

Round11-r'nonma_T4

Accuracy:	0.940605				
Precision:	0.998689	0.785714	0.991985	0.809561	0.943407
Recall/Ser	0.918072	0.929577	0.968694	0.753228	0.971055
Specificity	0.999826	0.997239	0.998872	0.979048	0.900144
F1_Score:	0.956686	0.851613	0.980201	0.780379	0.957031

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1524	18	6	7	105	1660
Dendritic_	0	132	5	0	5	142
Monocyte	1	17	1609	0	34	1661
NK_cells	1	0	2	1050	341	1394
T_cells	0	1	0	240	8085	8326
All	1526	168	1622	1297	8570	13183

Accuracy:	0.924585				
Precision:	0.955739	0.967742	0.88805	0.619592	0.984049
Recall/Ser	0.985669	0.111111	0.993433	0.901425	0.921413
Specificity	0.991737	0.999917	0.97372	0.959301	0.979134
F1_Score:	0.970473	0.199336	0.937791	0.734398	0.951701

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1857	0	9	0	18	1884
Dendritic_	19	30	213	4	4	270
Monocyte	8	0	2118	0	6	2132
NK_cells	2	0	2	759	79	842
T_cells	57	1	43	462	6601	7164
All	1943	31	2385	1225	6708	12292

Accuracy:	0.143708				
Precision:	0.695946	0.03151	1	0.945449	
Recall/Ser	2.04E-02	0.903139	0.017174	0.148692	
Specificity	0.998805	0.124452	1	0.973817	
F1_Score:	0.039688	0.060896	0.033767	0.25697	

Predicted	B_cells	Monocytes	NK_cells	T_cells	All
B_cells	206	9871	0	8	10085
Monocyte	5	2359	0	248	2612
NK_cells	0	7945	144	296	8385
T_cells	85	54689	0	9567	64341
All	296	74864	144	10119	85423

Accuracy:	0.846175				
Precision:	0.813066	0	0.853066	0.175099	0.949261
Recall/Ser	0.703807	0	0.733414	0.996764	0.89252
Specificity	0.982076	0.999944	0.971456	0.917562	0.889362
F1_Score:	0.754502	0	0.788728	0.297872	0.920016

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	1	204	79	245	1786
Monocyte	64	0	2421	465	351	3301
NK_cells	0	0	0	308	1	309
T_cells	225	0	213	907	11169	12514
All	1546	1	2838	1759	11766	17910

Round12-r'nonma_T4_afth

Accuracy:	0.938254				
Precision:	0.994167	0.795181	0.95283	0.787994	0.951312
Recall/Ser	0.924096	0.929577	0.972908	0.743902	0.966851
Specificity	0.999219	0.997393	0.993057	0.976334	0.915174
F1_Score:	0.957852	0.857143	0.962764	0.765314	0.959018

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1534	24	53	16	33	1660
Dendritic_	0	132	9	0	1	142
Monocyte	7	9	1616	0	29	1661
NK_cells	2	0	6	1037	349	1394
T_cells	0	1	12	263	8050	8326
All	1543	166	1696	1316	8462	13183

Accuracy:	0.898226				
Precision:	0.968421	0.5	0.867555	0.549801	0.954872
Recall/Ser	0.927813	0.011111	0.977017	0.819477	0.909687
Specificity	0.994523	0.99975	0.968701	0.950655	0.939938
F1_Score:	0.947682	0.021739	0.919038	0.658083	0.931732

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1748	0	111	15	10	1884
Dendritic_	4	3	120	2	141	270
Monocyte	31	0	2083	0	18	2132
NK_cells	4	0	9	690	139	842
T_cells	18	3	78	548	6517	7164
All	1805	6	2401	1255	6825	12292

Accuracy:	0.10208				
Precision:	0.642857	0	0.032306	0.992411	0.968488
Recall/Ser	7.23E-02	0	0.973201	0.14037	6.64E-02
Specificity	0.994624	0.999906	0.080533	0.999883	0.993407
F1_Score:	0.129958	0	0.062537	0.245951	0.124273

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	729	0	9354	0	2	10085
Monocyte	6	6	2542	3	55	2612
NK_cells	1	0	7125	1177	82	8385
T_cells	398	2	59663	6	4272	64341
All	1134	8	78684	1186	4411	85423

Accuracy:	0.806556				
Precision:	0.815704	0	0.853066	0.175699	0.925857
Recall/Ser	0.703807	0	0.733414	0.996764	0.848219
Specificity	0.977095	0.99993	0.961687	0.895863	0.889362
F1_Score:	0.755636	0	0.788728	0.298739	0.885339

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	1	204	79	245	1786
Monocyte	64	0	2421	465	351	3301
NK_cells	0	0	0	308	1	309
T_cells	220	0	213	901	7455	8789
All	1541	1	2838	1753	8052	14185

Round13-r'HLADR_48

Accuracy:	0.934916				
Precision:	0.997374	0.5	0.915501	0.785359	0.952848
Recall/Ser	0.91506	0.014085	0.984949	0.792683	0.968412
Specificity	0.999653	0.999847	0.986895	0.974383	0.917851
F1_Score:	0.954445	0.027397	0.948956	0.789004	0.960567

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1519	2	31	41	67	1660
Dendritic_	0	2	113	3	24	142
Monocyte	2	0	1636	0	23	1661
NK_cells	2	0	2	1105	285	1394
T_cells	0	0	5	258	8063	8326
All	1523	4	1787	1407	8462	13183

Accuracy:	0.934348				
Precision:	0.967658	0.966667	0.912688	0.641187	0.9795
Recall/Ser	0.984607	0.214815	0.99531	0.846793	0.940396
Specificity	0.994043	0.999834	0.98002	0.965153	0.972504
F1_Score:	0.976059	0.351515	0.95221	0.729785	0.95955

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1855	0	15	2	12	1884
Dendritic_	43	58	142	19	8	270
Monocyte	4	0	2122	2	4	2132
NK_cells	2	0	10	713	117	842
T_cells	13	2	36	376	6737	7164
All	1917	60	2325	1112	6878	12292

Accuracy:	0.195228				
Precision:	0.516402	0	0.034598	1	0.957132
Recall/Ser	4.84E-02	0	0.928025	0.006679	2.13E-01
Specificity	0.993934	0.999567	0.183224	1	0.970876
F1_Score:	0.088486	0	0.066709	0.013269	0.348546

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	488	0	9590	0	7	10085
Monocyte	17	36	2424	0	135	2612
NK_cells	0	1	7856	56	472	8385
T_cells	440	0	50192	0	13709	64341
All	945	37	70062	56	14323	85423

Accuracy:	0.80696				
Precision:	0.818359	0	0.851337	0.176	0.926663
Recall/Ser	0.703807	0	0.734092	0.996764	0.848219
Specificity	0.977411	0.999929	0.961687	0.895719	0.889678
F1_Score:	0.756773	0	0.788379	0.299174	0.885708

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	1	204	79	245	1786
Monocyte	59	0	2388	462	344	3253
NK_cells	0	0	0	308	1	309
T_cells	220	0	213	901	7455	8789
All	1536	1	2805	1750	8045	14137

Round14-r'HLADR_2397

Accuracy:	0.939923				
Precision:	0.996034	0.716578	0.970838	0.785509	0.954282
Recall/Ser	0.907831	0.943662	0.962071	0.785509	0.967692
Specificity	0.999479	0.995936	0.995834	0.974637	0.920527
F1_Score:	0.94989	0.81459	0.966435	0.785509	0.96094

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1507	19	39	33	62	1660
Dendritic_	0	134	7	0	1	142
Monocyte	2	33	1598	0	28	1661
NK_cells	2	0	2	1095	295	1394
T_cells	2	1	0	266	8057	8326
All	1513	187	1646	1394	8443	13183

Accuracy:	0.941019				
Precision:	0.962474	0.95122	0.969599	0.679918	0.962451
Recall/Ser	0.966561	0.577778	0.987336	0.792162	0.951703
Specificity	0.993178	0.999335	0.993504	0.972576	0.948128
F1_Score:	0.964513	0.718894	0.978387	0.731761	0.957047

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1821	1	6	0	56	1884
Dendritic_	48	156	43	0	23	270
Monocyte	8	6	2105	0	13	2132
NK_cells	1	0	0	667	174	842
T_cells	14	1	17	314	6818	7164
All	1892	164	2171	981	7084	12292

Accuracy:	0.749143				
Precision:	0.36071	0	1	0.99863	0.956204
Recall/Ser	9.98E-01	0	0.154288	0.608587	7.53E-01
Specificity	0.763293	0.98395	1	0.999909	0.894792
F1_Score:	0.529858	0	0.26733	0.75628	0.842301

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	10062	1	0	0	22	10085
Monocyte	76	1179	403	1	953	2612
NK_cells	1866	173	0	5103	1243	8385
T_cells	15891	18	0	6	48426	64341
All	27895	1371	403	5110	50644	85423

Accuracy:	0.839779				
Precision:	0.849324	0	0.667994	0.238206	0.966926
Recall/Ser	0.703807	0	0.98014	0.996764	0.848219
Specificity	0.977597	0.999915	0.961687	0.913831	0.913589
F1_Score:	0.769749	0	0.794508	0.384519	0.903691

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	1	204	79	245	1786
Monocyte	3	0	839	5	9	856
NK_cells	0	0	0	308	1	309
T_cells	220	0	213	901	7455	8789
All	1480	1	1256	1293	7710	11740

Round15-r'CD19_26

Accuracy:	0.944626				
Precision:	0.994325	0.832061	0.950839	0.768763	0.967062
Recall/Ser	0.95	0.767606	0.989765	0.815638	0.959164
Specificity	0.999219	0.998313	0.992623	0.97099	0.943998
F1_Score:	0.971657	0.798535	0.969912	0.791507	0.963097

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1577	15	50	8	10	1660
Dendritic_	0	109	32	0	1	142
Monocyte	4	5	1644	0	8	1661
NK_cells	2	1	1	1137	253	1394
T_cells	3	1	2	334	7986	8326
All	1586	131	1729	1479	8258	13183

Accuracy:	0.9363				
Precision:	0.966475	0.965753	0.960927	0.611529	0.976166
Recall/Ser	0.979299	0.522222	0.992026	0.869359	0.931882
Specificity	0.993851	0.999584	0.991535	0.959389	0.968214
F1_Score:	0.972845	0.677885	0.976229	0.717999	0.95351

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1845	0	5	0	34	1884
Dendritic_	58	141	59	4	8	270
Monocyte	2	3	2115	0	12	2132
NK_cells	0	1	0	732	109	842
T_cells	4	1	22	461	6676	7164
All	1909	146	2201	1197	6839	12292

Accuracy:	0.879365				
Precision:	0.609836	0	0.995839	0.999627	0.946481
Recall/Ser	9.82E-01	0	0.274885	0.638521	9.19E-01
Specificity	0.915859	0.992777	0.999964	0.999974	0.841381
F1_Score:	0.752544	0	0.430843	0.779274	0.932607

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	9908	3	0	0	174	10085
Monocyte	114	583	718	0	1197	2612
NK_cells	1046	10	2	5354	1973	8385
T_cells	5179	21	1	2	59138	64341
All	16247	617	721	5356	62482	85423

Accuracy:	0.84096			
Precision:	0.848505	0.671737	0.238206	0.968182
Recall/Ser	0.709659	0.98014	0.996764	0.848219
Specificity	0.977597	0.96224	0.913634	0.916239
F1_Score:	0.772896	0.79715	0.384519	0.904239

Predicted	B_cells	Monocytes	NK_cells	T_cells	All
B_cells	1249	197	79	235	1760
Monocyte	3	839	5	9	856
NK_cells	0	0	308	1	309
T_cells	220	213	901	7455	8789
All	1472	1249	1293	7700	11714

Round16-r'CD19_1760

Accuracy:	0.942426				
Precision:	0.998692	0.794118	0.978274	0.783297	0.954373
Recall/Ser	0.91988	0.950704	0.975918	0.780488	0.967211
Specificity	0.999826	0.997316	0.996876	0.974468	0.920733
F1_Score:	0.957667	0.865385	0.977095	0.78189	0.960749

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1527	22	26	29	56	1660
Dendritic_	0	135	7	0	0	142
Monocyte	0	13	1621	0	27	1661
NK_cells	2	0	2	1088	302	1394
T_cells	0	0	1	272	8053	8326
All	1529	170	1657	1389	8438	13183

Accuracy:	0.904653				
Precision:	0.927228	0	0.939019	0.492932	0.99044
Recall/Ser	0.994161	0	0.996717	0.952494	0.882189
Specificity	0.985876	0.999917	0.986417	0.927948	0.988105
F1_Score:	0.959529	0	0.967008	0.649656	0.933186

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1873	0	2	1	8	1884
Dendritic_	125	0	122	12	11	270
Monocyte	2	0	2125	0	5	2132
NK_cells	2	0	1	802	37	842
T_cells	18	1	13	812	6320	7164
All	2020	1	2263	1627	6381	12292

Accuracy:	0.877328				
Precision:	0.996643	0	1	0.998844	0.861523
Recall/Ser	5.00E-01	0	0.148545	0.618485	1.00E+00
Specificity	0.999774	0.99863	1	0.999922	0.509582
F1_Score:	0.666315	0	0.258667	0.763939	0.925491

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	5047	1	0	0	5037	10085
Monocyte	14	104	388	3	2103	2612
NK_cells	0	0	0	5186	3199	8385
T_cells	3	12	0	3	64323	64341
All	5064	117	388	5192	74662	85423

Accuracy:	0.864175			
Precision:	0	0.797529	0.253707	0.99866
Recall/Ser	0	0.98014	0.996764	0.848219
Specificity	0.977597	0.976588	0.906065	0.991416
F1_Score:	0	0.879455	0.404465	0.917313

Predicted	B_cells	Monocytes	NK_cells	T_cells	All
Monocyte	3	839	5	9	856
NK_cells	0	0	308	1	309
T_cells	220	213	901	7455	8789
All	223	1052	1214	7465	9954

Round17-r'CD8_5662

Accuracy:	0.94614276					
Precision:	0.99806076	0.81437126	0.99323493	0.79407407	0.95443306	
Recall/Sensitivity:	0.93012048	0.95774648	0.97230584	0.76901004	0.97357675	
Specificity:	0.99973965	0.99762288	0.9990453	0.9764187	0.92032119	
F1_Score:	0.96289367	0.8802589	0.98265896	0.78134111	0.96390986	
Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1544	20	5	60	31	1660
Dendritic_cells	0	136	5	0	1	142
Monocytes	1	9	1615	0	36	1661
NK_cells	2	0	1	1072	319	1394
T_cells	0	2	0	218	8106	8326
All	1547	167	1626	1350	8493	13183

Accuracy:	0.917344614					
Precision:	0.93098312	0	0.92384682	0.55555556	0.98828897	
Recall/Sensitivity:	0.99522293	0	0.99577861	0.9263658	0.90703518	
Specificity:	0.98664489	0.99991682	0.98277559	0.94550218	0.9849844	
F1_Score:	0.96203181	0	0.95846501	0.69456812	0.94592037	
Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1875	0	6	0	3	1884
Dendritic_cells	103	0	152	0	15	270
Monocytes	6	0	2123	0	3	2132
NK_cells	6	0	0	780	56	842
T_cells	24	1	17	624	6498	7164
All	2014	1	2298	1404	6575	12292

Accuracy:	0.982920291					
Precision:	0.9769316	0	0.84932995	0.98514663	0.99214461	
Recall/Sensitivity:	9.62E-01	0	0.89777948	0.92546213	9.97E-01	
Specificity:	0.99696037	0.99778748	0.99497651	0.99848127	0.97590361	
F1_Score:	0.96921847	0	0.87288293	0.95437216	0.99466708	
Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	9698	28	356	1	2	10085
Monocytes	202	19	2345	3	43	2612
NK_cells	0	135	27	7760	463	8385
T_cells	27	7	33	113	64161	64341
All	9927	189	2761	7877	64669	85423

Accuracy:	0.935228332				
Precision:	0	0.99762188	0.53940455	0.99652416	
Recall/Sensitivity:	0	0.98014019	0.99676375	0.91685321	
Specificity:	0.99930103	0.99941793	0.93396937	0.99141631	
F1_Score:	0	0.98880377	0.7	0.95502998	
Predicted	B_cells	Monocytes	NK_cells	T_cells	All
Monocytes	3	839	5	9	856
NK_cells	0	0	308	1	309
T_cells	0	2	258	2867	3127
All	3	841	571	2877	4292

- Subtype classification performance (1-Sensitivity) during group comparison.

TestWithBroadS1		R1	R5	R7	R8	R12	R17
	Bn	0.082121	0.079555	0.083832	0.084688	0.071856	0.063302
	Bm	0.103870	0.087576	0.091650	0.095723	0.085540	0.085540
	DC	0.119718	0.260563	0.084507	0.070423	0.070423	0.042254
	M14	0.063341	0.018211	0.015835	0.053048	0.024545	0.026920
	M16	0.080402	0.022613	0.027638	0.067839	0.035176	0.030151
	NK	0.174319	0.256098	0.190100	0.210904	0.256098	0.230990
	aTreg	0.001086	0.001086	0.001086	0.002172	0.001086	0.001086
	nonT	0.549296	0.448357	0.485915	0.448357	0.427230	0.422535
	rTreg	0.003731	0.004664	0.002799	0.003731	0.003731	0.000000
	T4em	0.005128	0.007179	0.009231	0.009231	0.004103	0.000000
	T4naive	0.002646	0.001764	0.000882	0.002646	0.003527	0.000882
	T8em	0.075655	0.039767	0.068865	0.053346	0.050436	0.028128
	T8naive	0.000749	0.000000	0.000000	0.000000	0.000000	0.000000
	Tncl	0.023760	0.016073	0.023061	0.020266	0.020266	0.006289

TestWithBroadS2		R1	R5	R7	R8	R12	R17
	BC	0.042463	0.126327	0.098726	0.080679	0.072187	0.004777
	DC	0.925743	0.698020	0.559406	0.767327	0.985149	1.000000
	pDC	0.985294	0.941176	0.897059	1.000000	1.000000	1.000000
	M14	0.027640	0.064124	0.004422	0.018242	0.021559	0.003870
	M16	0.024768	0.024768	0.021672	0.018576	0.030960	0.006192
	NK	0.077197	0.241093	0.210214	0.188836	0.180523	0.073634
	T4	0.036982	0.013314	0.009763	0.015680	0.019527	0.021006
	T8	0.180233	0.076903	0.073203	0.116015	0.153541	0.157241

TestWith10x		R1	R5	R7	R8	R12	R17
	BC	0.994348	0.904908	0.984432	0.964601	0.927714	0.038374
	M14	0.002297	0.081164	0.241960	0.083461	0.026799	0.102221
	NK	0.999761	0.969231	0.979249	0.999404	0.859630	0.074538
	CD45RA+CD25-T4naive	0.997233	0.958393	0.885390	0.929478	0.971371	0.004199
	T4	0.977704	0.945242	0.835191	0.910818	0.957906	0.002140
	CD45RA+T8naive	0.998159	0.948297	0.926378	0.958086	0.979670	0.000920
	T8	0.963855	0.802723	0.687335	0.845626	0.917034	0.007934
	CD45RO+T4mem	0.942293	0.835290	0.673807	0.812696	0.898963	0.000293
	CD4+CD25+Treg	0.889019	0.808536	0.592614	0.764981	0.865829	0.001656

TestWithGEO		R1-17
	empty_cells	1.000000
	tumor_ascites_DC	1.000000
	tonsil_DC	1.000000
	T8_methanol_SSC	0.298127
	donor1_ IL-10-producing_Foxp3-_T4	0.004812
	donor2_ IL-10-producing_Foxp3-_T4	0.006835
	nonmalignant_P5_CD3+CD5intSSCint_T4	0.006910
	nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy	0.002953
	HLA-DR	0.312500
	HLA-DR_control	0.353776
	CD19	0.692308
	CD19_control	0.290341
	CD8	0.189686

	R1-17
M14_d1	0.011765
M14_d2	0.027842
NK	0.003236
T4	0.000000
T8	0.016129
iNKT	0.113846
MAIT	0.052356
Vd1	0.454225
Vd2	0.215686
T4	0.016580
CCR5+CD69-T4	0.020690

- F1-score of each cell type during 17 rounds of four-supersets-swapping external cross-validation experiments.

TestWith-Source-BroadS1	Round1-A	Round2-A	Round3-A	Round4-A	Round5-r'	Round6-r'	Round7-r'	Round8-r'	Round9-r'
B_cells	0.953371	0.965217	0.951872	0.952411	0.955486	0.949325	0.953788	0.952201	0.946203
Dendritic_cells	0.692521	0.782609	0.778816	0.744615	0.769231	0	0.875421	0.73743	0.8
Monocytes	0.955583	0.969805	0.975684	0.969254	0.96362	0.953901	0.975756	0.958117	0.961376
NK_cells	0.787817	0.775144	0.762509	0.77169	0.769288	0.666424	0.775412	0.783197	0.785765
T_cells	0.955901	0.960139	0.951129	0.957128	0.957566	0.943473	0.961948	0.959279	0.961612

Round10-r	Round11-r	Round12-r	Round13-r	Round14-r	Round15-r	Round16-r	Round17-r
0.961144	0.956686	0.957852	0.954445	0.94989	0.971657	0.957667	0.962894
0.786787	0.851613	0.857143	0.027397	0.81459	0.798535	0.865385	0.880259
0.976829	0.980201	0.962764	0.948956	0.966435	0.969912	0.977095	0.982659
0.784938	0.780379	0.765314	0.789004	0.785509	0.791507	0.78189	0.781341
0.960534	0.957031	0.959018	0.960567	0.96094	0.963097	0.960749	0.96391

TestWith-Source-BroadS2	Round1-A	Round2-A	Round3-A	Round4-A	Round5-r'	Round6-r'	Round7-r'	Round8-r'	Round9-r'
B_cells	0.967033	0.960064	0.930307	0.954989	0.909895	0.934409	0.94072	0.941304	0.944004
Dendritic_cells	0.107744	0.656652	0.236422	0.347305	0.360111	0.535	0.518919	0.291022	0.253165
Monocytes	0.948548	0.95114	0.886801	0.955355	0.919414	0.95932	0.969766	0.92284	0.916923
NK_cells	0.642149	0.698885	0.657213	0.60274	0.712375	0.710994	0.741774	0.702675	0.663239
T_cells	0.923781	0.958062	0.931895	0.930007	0.94551	0.957517	0.948979	0.946316	0.943169

Round10-r	Round11-r	Round12-r	Round13-r	Round14-r	Round15-r	Round16-r	Round17-r
0.98344	0.970473	0.947682	0.976059	0.964513	0.972845	0.959529	0.962032
0	0.199336	0.021739	0.351515	0.718894	0.677885	0	0
0.948931	0.937791	0.919038	0.95221	0.978387	0.976229	0.967008	0.958465
0.536439	0.734398	0.658083	0.729785	0.731761	0.717999	0.649656	0.694568
0.898018	0.951701	0.931732	0.95955	0.957047	0.95351	0.933186	0.94592

TestWith-Source-10x	Round1-A	Round2-A	Round3-A	Round4-A	Round5-r'	Round6-r'	Round7-r'	Round8-r'	Round9-r'
B_cells	0.011212	0.031212	0.042437	0.06844	0.159753	0.027722	0.030129	0.067929	0.18847
Dendritic_cells	0	0	0	0	0	0	0	0	0
Monocytes	0.060978	0.055378	0.060186	0.060754	0.061488	0.055985	0.055375	0.060595	0.065979
NK_cells	0.000477	0.185418	0.000477	0.000477	0.059701	0.000477	0.040659	0.001192	0.083838
T_cells	0.071807	0.248987	0.054488	0.118591	0.203362	0.107496	0.364322	0.223344	0.370334

Round10-r	Round11-r	Round12-r	Round13-r	Round14-r	Round15-r	Round16-r	Round17-r
0.089664	0.039688	0.129958	0.088486	0.529858	0.752544	0.666315	0.969218
0	0	0	0	0	0	0	0
0.05744	0.060896	0.062537	0.066709	0.26733	0.430843	0.258667	0.872883
0.002144	0.033767	0.245951	0.013269	0.75628	0.779274	0.763939	0.954372
0.152344	0.25697	0.124273	0.348546	0.842301	0.932607	0.925491	0.994667

TestWith-Source-GEODB	Round1-A	Round2-A	Round3-A	Round4-A	Round5-r'	Round6-r'	Round7-r'	Round8-r'	Round9-r'
B_cells	0.687073	0.688013	0.688578	0.688767	0.688956	0.70777	0.714814	0.754049	0.754049
Dendritic_cells	0.004532	0.002282	0.000916	0.000459	0	0			
Monocytes	0.483378	0.48362	0.483765	0.483813	0.483861	0.558542	0.773606	0.785911	0.785911
NK_cells	0.180328	0.180592	0.180751	0.180804	0.180857	0.189073	0.190712	0.294174	0.295019
T_cells	0.922434	0.922531	0.922589	0.922608	0.922627	0.923206	0.928366	0.94957	0.946342

Round10-r	Round11-r	Round12-r	Round13-r	Round14-r	Round15-r	Round16-r	Round17-r
0.754275	0.754502	0.755636	0.756773	0.769749	0.772896		
0.785911	0.788728	0.788728	0.788379	0.794508	0.79715	0.879455	0.988804
0.296724	0.297872	0.298739	0.299174	0.384519	0.384519	0.404465	0.7
0.94061	0.920016	0.885339	0.885708	0.903691	0.904239	0.917313	0.95503

- Split confusion matrix results of group comparison.

SplitConfusionMatrix-R1(10EC*5)

(R1 included ALL groups: all non-representative GEO datasets, and 10EC*5 in GEO 5-classes.)

Train: 10xall(Clean)+GEOall+BroadS2all(Clean)+++10EC-x-five

Test: BroadS1

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
1	10x (Clean)	BC	10085	85423	√	
		M14	2612		√	
		NK	8385		√	
		CD45RA+CD25-T4naive	10479		√	
		T4	11213		√	
		CD45RA+T8naive	11953		√	
		T8	10209		√	
		CD45RO+T4mem	10224		√	
		CD4+CD25+Treg	10263		√	
	GEO (ALL+10EC*5)	M14_d1	425	34700	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T4	435		√	
		tumor_ascites_DC	1613		√	
		tonsil_DC	2739		√	
		T8_methanol_SSC	4753		√	
		donor1_IL-10-producing_Foxp3-_T4	1247		√	
		donor2_IL-10-producing_Foxp3-_T4	1902		√	
		nonmalignant_P5_CD3+CD5intSSCir	4486		√	
		nonmalignant_P5_CD3+CD5intSSCir	3725		√	
		HLA-DR	48		√	
		HLA-DR_control	2397		√	
		CD19	26		√	
		CD19_control	1760		√	
		CD8	5662		√	
		10-empty-cells-in-BC	10		√	
		10-empty-cells-in-DC	10		√	
		10-empty-cells-in-MC	10		√	
		10-empty-cells-in-NK	10		√	
		10-empty-cells-in-TC	10		√	
	BroadS1	Bn	1169	13183		√
		Bm	491			√
		DC	142			√
		M14	1263			√
		M16	398			√
		NK	1394			√
		aTreg	921			√
		nonT	426			√
		rTreg	1072			√
		T4em	975			√
		T4naive	1134			√
		T8em	1031			√
		T8naive	1336			√
		Tncl	1431			√
	BroadS2 (Clean)	BC	1884	12292	√	
		DC	202		√	
		pDC	68		√	
		M14	1809		√	
		M16	323		√	
		NK	842		√	
		T4	3380		√	
		T8	3784		√	

Accuracy:	0.933323219				
Precision:	0.9993395	0.57077626	0.97975965	0.75327225	0.9550414
Recall/Sensitivi	0.91144578	0.88028169	0.93257074	0.82568149	0.956762
Specificity:	0.99991322	0.99279196	0.9972227	0.96802104	0.9227919
F1_Score:	0.95337114	0.69252078	0.95558297	0.78781656	0.9559009

Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1513	25	16	18	88	1660
Dendritic_cells	0	125	14	0	3	142
Monocytes	0	68	1549	0	44	1661
NK_cells	1	0	2	1151	240	1394
T_cells	0	1	0	359	7966	8326
All	1514	219	1581	1528	8341	13183

True/ Predicted					BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	Bn	Bn_aTreg	BT580	Bn_aTreg_BT580__BC	4					1169	0.0821	1660
			BT860	Bn_aTreg_BT860__BC	6							
			NY860	Bn_aTreg_NY860__BC	3							
		Bn_nonT	BT580	Bn_nonT_BT580__BC	234							
				Bn_nonT_BT580__DC		2						
				Bn_nonT_BT580__MC			1					
				Bn_nonT_BT580__NK				2				
				Bn_nonT_BT580__TC					8			
			BT860	Bn_nonT_BT860__BC	511							
				Bn_nonT_BT860__DC		6						
				Bn_nonT_BT860__MC			7					
				Bn_nonT_BT860__NK				3				
				Bn_nonT_BT860__TC					26			
			NY580	Bn_nonT_NY580__BC	148							
				Bn_nonT_NY580__DC		2						
				Bn_nonT_NY580__MC			2					
				Bn_nonT_NY580__NK				1				
				Bn_nonT_NY580__TC					11			
			NY860	Bn_nonT_NY860__BC	165							
				Bn_nonT_NY860__DC		6						
				Bn_nonT_NY860__NK				6				
				Bn_nonT_NY860__TC					13			
		Bn_T4em	BT860	Bn_T4em_BT860__BC	1							
		Bn_Tncl	BT860	Bn_Tncl_BT860__BC	1							
	Bm	Bm_aTreg	BT860	Bm_aTreg_BT860__BC	6					491	0.1039	
			NY580	Bm_aTreg_NY580__BC	1							
			NY860	Bm_aTreg_NY860__BC	2							
		Bm_nonT	BT580	Bm_nonT_BT580__BC	86							
				Bm_nonT_BT580__MC			1					
				Bm_nonT_BT580__TC					1			
			BT860	Bm_nonT_BT860__BC	206							
				Bm_nonT_BT860__DC		2						
				Bm_nonT_BT860__MC			2					
				Bm_nonT_BT860__NK				1				
				Bm_nonT_BT860__TC					12			
			NY580	Bm_nonT_NY580__BC	58							
				Bm_nonT_NY580__DC		1						
				Bm_nonT_NY580__NK				1				
				Bm_nonT_NY580__TC					3			
			NY860	Bm_nonT_NY860__BC	81							
				Bm_nonT_NY860__DC		6						
				Bm_nonT_NY860__MC			3					
				Bm_nonT_NY860__NK				4				
				Bm_nonT_NY860__TC					14			
Dendritic_cells	DC	DC_aTreg	BT860	DC_aTreg_BT860__DC		1				142	0.1197	142
			NY580	DC_aTreg_NY580__DC		1						
		DC_nonT	BT580	DC_nonT_BT580__DC		46						
				DC_nonT_BT580__MC			7					
				DC_nonT_BT580__TC					1			
			BT860	DC_nonT_BT860__DC		17						
				DC_nonT_BT860__MC			1					
				DC_nonT_BT860__TC					1			
			NY580	DC_nonT_NY580__DC		43						
				DC_nonT_NY580__MC			3					
			NY860	DC_nonT_NY860__DC		17						
				DC_nonT_NY860__MC			3					
				DC_nonT_NY860__TC					1			
Monocytes	M14	M14_aTreg	BT580	M14_aTreg_BT580__MC			1			1263	0.0633	1661
			BT860	M14_aTreg_BT860__MC			4					
			NY580	M14_aTreg_NY580__MC			2					
			NY860	M14_aTreg_NY860__MC			2					
		M14_nonT	BT580	M14_nonT_BT580__DC		19						
				M14_nonT_BT580__MC			215					
				M14_nonT_BT580__TC					4			
			BT860	M14_nonT_BT860__DC		15						
				M14_nonT_BT860__MC			315					
				M14_nonT_BT860__TC					8			
			NY580	M14_nonT_NY580__DC		8						
				M14_nonT_NY580__MC			328					
				M14_nonT_NY580__TC					5			
			NY860	M14_nonT_NY860__DC		8						
				M14_nonT_NY860__MC			314					
				M14_nonT_NY860__TC					13			
		M14_rTreg	NY580	M14_rTreg_NY580__MC			1					
		M14_Tncl	BT580	M14_Tncl_BT580__MC			1					
	M16	M16_aTreg	BT580	M16_aTreg_BT580__MC			4			398	0.0804	
			BT860	M16_aTreg_BT860__MC			5					
			NY580	M16_aTreg_NY580__MC			7					
			NY860	M16_aTreg_NY860__MC			7					
		M16_nonT	BT580	M16_nonT_BT580__DC		2						
				M16_nonT_BT580__MC			57					
			BT860	M16_nonT_BT860__DC		6						
				M16_nonT_BT860__MC			92					
				M16_nonT_BT860__TC					9			
			NY580	M16_nonT_NY580__DC		3						
				M16_nonT_NY580__MC			75					
				M16_nonT_NY580__TC					3			
			NY860	M16_nonT_NY860__DC		7						
				M16_nonT_NY860__MC			117					
				M16_nonT_NY860__TC					2			
		M16_T8em	BT580	M16_T8em_BT580__MC			1					
		M16_T8em	NY860	M16_T8em_NY860__MC			1					

NK_cells	NK	NK_aTreg	BT580	NK_aTreg_BT580___TC					2	1394	0.1743	1394
			NY580	NK_aTreg_NY580___TC					3			
			NY860	NK_aTreg_NY860___TC					1			
		NK_nonT	BT580	NK_nonT_BT580___MC			1					
				NK_nonT_BT580___NK				242				
				NK_nonT_BT580___TC					11			
			BT860	NK_nonT_BT860___BC	1							
				NK_nonT_BT860___NK				374				
				NK_nonT_BT860___TC					50			
			NY580	NK_nonT_NY580___MC			1					
				NK_nonT_NY580___NK				180				
				NK_nonT_NY580___TC					7			
			NY860	NK_nonT_NY860___NK				240				
				NK_nonT_NY860___TC					24			
		NK_T4em	NY860	NK_T4em_NY860___NK				1				
		NK_T4naive	NY860	NK_T4naive_NY860___TC					1			
		NK_T8em	BT580	NK_T8em_BT580___NK				20				
				NK_T8em_BT580___TC					20			
			BT860	NK_T8em_BT860___NK				37				
				NK_T8em_BT860___TC					49			
			NY580	NK_T8em_NY580___NK				13				
				NK_T8em_NY580___TC					5			
			NY860	NK_T8em_NY860___NK				33				
				NK_T8em_NY860___TC					35			
		NK_Tncl	BT580	NK_Tncl_BT580___NK				2				
				NK_Tncl_BT580___TC					8			
			BT860	NK_Tncl_BT860___NK				3				
				NK_Tncl_BT860___TC					7			
			NY580	NK_Tncl_NY580___NK				2				
				NK_Tncl_NY580___TC					9			
			NY860	NK_Tncl_NY860___NK				4				
				NK_Tncl_NY860___TC					8			
T_cells	aTreg	T_aTreg	BT580	T_aTreg_BT580___TC					241	921	0.0011	8326
			BT860	T_aTreg_BT860___TC					243			
			NY580	T_aTreg_NY580___NK				1				
				T_aTreg_NY580___TC					221			
			NY860	T_aTreg_NY860___TC					215			
	nonT	T_nonT	BT580	T_nonT_BT580___NK				56		426	0.5493	
				T_nonT_BT580___TC					40			
			BT860	T_nonT_BT860___NK				61				
				T_nonT_BT860___TC					73			
			NY580	T_nonT_NY580___NK				59				
				T_nonT_NY580___TC					26			
			NY860	T_nonT_NY860___NK				58				
				T_nonT_NY860___TC					53			
	rTreg	T_rTreg	BT580	T_rTreg_BT580___NK				2		1072	0.0037	
				T_rTreg_BT580___TC					311			
			BT860	T_rTreg_BT860___NK				1				
				T_rTreg_BT860___TC					233			
			NY580	T_rTreg_NY580___TC					337			
			NY860	T_rTreg_NY860___NK				1				
				T_rTreg_NY860___TC					187			
	T4em	T_T4em	BT580	T_T4em_BT580___NK				2		975	0.0051	
				T_T4em_BT580___TC					328			
			BT860	T_T4em_BT860___NK				2				
				T_T4em_BT860___TC					257			
			NY580	T_T4em_NY580___NK				1				
				T_T4em_NY580___TC					253			
			NY860	T_T4em_NY860___TC					132			
	T4naive	T_T4naive	BT580	T_T4naive_BT580___DC		1				1134	0.0026	
				T_T4naive_BT580___NK				1				
				T_T4naive_BT580___TC					480			
			BT860	T_T4naive_BT860___TC					265			
			NY580	T_T4naive_NY580___NK				1				
				T_T4naive_NY580___TC					290			
			NY860	T_T4naive_NY860___TC					96			
	T8em	T_T8em	BT580	T_T8em_BT580___NK				20		1031	0.0757	
				T_T8em_BT580___TC					246			
			BT860	T_T8em_BT860___NK				21				
				T_T8em_BT860___TC					283			
			NY580	T_T8em_NY580___NK				19				
				T_T8em_NY580___TC					247			
			NY860	T_T8em_NY860___NK				18				
				T_T8em_NY860___TC					177			
	T8naive	T_T8naive	BT580	T_T8naive_BT580___TC					318	1336	0.0007	
			BT860	T_T8naive_BT860___TC					486			
			NY580	T_T8naive_NY580___NK				1				
				T_T8naive_NY580___TC					255			
			NY860	T_T8naive_NY860___TC					276			
	Tncl	T_Tncl	BT580	T_Tncl_BT580___NK				8		1431	0.0238	
				T_Tncl_BT580___TC					193			
			BT860	T_Tncl_BT860___NK				5				
				T_Tncl_BT860___TC					361			
			NY580	T_Tncl_NY580___NK				8				
				T_Tncl_NY580___TC					371			
			NY860	T_Tncl_NY860___NK				13				
				T_Tncl_NY860___TC					472			
All (predicted)					**1514**	**219**	**1581**	**1528**	**8341**	**13183**		**13183**

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
2	10x (Clean)	BC	10085	85423	√	
		M14	2612		√	
		NK	8385		√	
		CD45RA+CD2	10479		√	
		T4	11213		√	
		CD45RA+T8n	11953		√	
		T8	10209		√	
		CD45RO+T4m	10224		√	
		CD4+CD25+T	10263		√	
	GEO (ALL+10EC*5)	M14_d1	425	34700	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T	435		√	
		tumor_ascite	1613		√	
		tonsil_DC	2739		√	
		T8_methanol	4753		√	
		donor1_ IL-1(	1247		√	
		donor2_ IL-1(	1902		√	
		nonmalignan	4486		√	
		nonmalignan	3725		√	
		HLA-DR	48		√	
		HLA-DR_cont	2397		√	
		CD19	26		√	
		CD19_contro	1760		√	
		CD8	5662		√	
		10-empty-ce	10		√	
		10-empty-ce	10		√	
		10-empty-ce	10		√	
		10-empty-ce	10		√	
		10-empty-ce	10		√	
	BroadS1	Bn	1169	13183	√	
		Bm	491		√	
		DC	142		√	
		M14	1263		√	
		M16	398		√	
		NK	1394		√	
		aTreg	921		√	
		nonT	426		√	
		rTreg	1072		√	
		T4em	975		√	
		T4naive	1134		√	
		T8em	1031		√	
		T8naive	1336		√	
		Tncl	1431		√	
	BroadS2 (Clean)	BC	1884	12292		√
		DC	202			√
		pDC	68			√
		M14	1809			√
		M16	323			√
		NK	842			√
		T4	3380			√
		T8	3784			√

Accuracy:	0.897169				
Precision:	0.976719	0.592593	0.9254797	0.49239544	0.963328
Recall/Ser	0.957537	0.059259	0.9727955	0.92280285	0.887353
Specificity	0.995869	0.999085	0.98356299	0.93004367	0.952808
F1_Score:	0.967033	0.107744	0.94854791	0.64214876	0.923781

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1804	0	32	0	48	1884
Dendritic_	15	16	116	11	112	270
Monocyte	17	10	2074	11	20	2132
NK_cells	0	0	3	777	62	842
T_cells	11	1	16	779	6357	7164
All	1847	27	2241	1578	6599	12292

True/ Predicted						BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	BC	pbmc1	v2	A	pbmc1_v2_A_BC___BC	271					1884	0.0425	1884
					pbmc1_v2_A_BC___MC			5					
					pbmc1_v2_A_BC___TC					12			
				B	pbmc1_v2_B_BC___BC	356							
					pbmc1_v2_B_BC___MC			8					
					pbmc1_v2_B_BC___TC					24			
			v3		pbmc1_v3_BC___BC	324							
					pbmc1_v3_BC___MC			15					
					pbmc1_v3_BC___TC					7			
		pbmc2	v2		pbmc2_V2_BC___BC	853							
					pbmc2_V2_BC___MC			4					
					pbmc2_V2_BC___TC					5			
Dendritic_cells	DC	pbmc1	v2	A	pbmc1_v2_A_DC___BC	4					202	0.9257	270
					pbmc1_v2_A_DC___DC		2						
					pbmc1_v2_A_DC___MC			29					
					pbmc1_v2_A_DC___TC					20			
				B	pbmc1_v2_B_DC___MC			18					
					pbmc1_v2_B_DC___TC					15			
			v3		pbmc1_v3_DC___MC			10					
					pbmc1_v3_DC___NK				4				
					pbmc1_v3_DC___TC					24			
		pbmc2	v2		pbmc2_V2_DC___BC	4							
					pbmc2_V2_DC___DC		13						
					pbmc2_V2_DC___MC			15					
					pbmc2_V2_DC___TC					44			
	pDC	pbmc1	v2	A	pbmc1_v2_A_pDC___BC	1					68	0.9853	
					pbmc1_v2_A_pDC___MC			21					
					pbmc1_v2_A_pDC___NK				4				
				B	pbmc1_v2_B_pDC___MC			7					
					pbmc1_v2_B_pDC___NK				2				
					pbmc1_v2_B_pDC___TC					3			
		pbmc2	V2		pbmc2_V2_pDC___BC	6							
					pbmc2_V2_pDC___DC		1						
					pbmc2_V2_pDC___MC			16					
					pbmc2_V2_pDC___NK				1				
					pbmc2_V2_pDC___TC					6			
Monocytes	M14	pbmc1	v2	A	pbmc1_v2_A_M14___BC	9					1809	0.0276	2132
					pbmc1_v2_A_M14___DC		6						
					pbmc1_v2_A_M14___MC			609					
					pbmc1_v2_A_M14___NK				6				
					pbmc1_v2_A_M14___TC					10			
				B	pbmc1_v2_B_M14___BC	3							
					pbmc1_v2_B_M14___MC			373					
					pbmc1_v2_B_M14___NK				2				
					pbmc1_v2_B_M14___TC					1			
			v3		pbmc1_v3_M14___BC	1							
					pbmc1_v3_M14___DC		1						
					pbmc1_v3_M14___MC			350					
					pbmc1_v3_M14___NK				1				
					pbmc1_v3_M14___TC					1			
		pbmc2	V2		pbmc2_V2_M14___BC	4							
					pbmc2_V2_M14___DC		2						
					pbmc2_V2_M14___MC			427					
					pbmc2_V2_M14___TC					3			
	M16	pbmc1	v2	A	pbmc1_v2_A_M16___DC		1				323	0.0248	
					pbmc1_v2_A_M16___MC			94					
					pbmc1_v2_A_M16___NK				2				
					pbmc1_v2_A_M16___TC					5			
				B	pbmc1_v2_B_M16___MC			73					
			v3		pbmc1_v3_M16___MC			98					
		pbmc2	V2		pbmc2_V2_M16___MC			50					
NK_cells	NK	pbmc1	v2	A	pbmc1_v2_A_NK___NK				157		842	0.0772	842
					pbmc1_v2_A_NK___TC					9			
				B	pbmc1_v2_B_NK___MC			3					
					pbmc1_v2_B_NK___NK				230				
					pbmc1_v2_B_NK___TC					30			
			v3		pbmc1_v3_NK___NK				177				
					pbmc1_v3_NK___TC					17			
		pbmc2	V2		pbmc2_V2_NK___NK				213				
					pbmc2_V2_NK___TC					6			
T_cells	T4	pbmc1	v2	A	pbmc1_v2_A_T4___BC	2					3380	0.0370	7164
					pbmc1_v2_A_T4___NK				25				
					pbmc1_v2_A_T4___TC					523			
				B	pbmc1_v2_B_T4___BC	1							
					pbmc1_v2_B_T4___MC			1					
					pbmc1_v2_B_T4___NK				31				
					pbmc1_v2_B_T4___TC					875			
			v3		pbmc1_v3_T4___NK				48				
					pbmc1_v3_T4___TC					912			
		pbmc2	V2		pbmc2_V2_T4___BC	5							
					pbmc2_V2_T4___DC		1						
					pbmc2_V2_T4___MC			4					
					pbmc2_V2_T4___NK				7				
					pbmc2_V2_T4___TC					945			
	T8	pbmc1	v2	A	pbmc1_v2_A_T8___BC	1					3784	0.1802	
					pbmc1_v2_A_T8___MC			5					
					pbmc1_v2_A_T8___NK				250				
					pbmc1_v2_A_T8___TC					918			
				B	pbmc1_v2_B_T8___MC			2					
					pbmc1_v2_B_T8___NK				185				
					pbmc1_v2_B_T8___TC					767			
			v3		pbmc1_v3_T8___NK				148				
					pbmc1_v3_T8___TC					814			
		pbmc2	V2		pbmc2_V2_T8___BC	2							
					pbmc2_V2_T8___MC			4					
					pbmc2_V2_T8___NK				85				
					pbmc2_V2_T8___TC					603			
All (predicted)						1847	27	2241	1578	6599	12292		12292

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
3	10x (Clean)	BC	10085	85423		√
		M14	2612			√
		NK	8385			√
		CD45RA+CD25-T4naive	10479			√
		T4	11213			√
		CD45RA+T8naive	11953			√
		T8	10209			√
		CD45RO+T4mem	10224			√
		CD4+CD25+Treg	10263			√
	GEO (ALL+10EC*5)	M14_d1	425	34700	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T4	435		√	
		tumor_ascites_DC	1613		√	
		tonsil_DC	2739		√	
		T8_methanol_SSC	4753		√	
		donor1_ IL-10-producing_Foxp3-_T4	1247		√	
		donor2_ IL-10-producing_Foxp3-_T4	1902		√	
		nonmalignant_P5_CD3+CD5intSSCint_T4	4486		√	
		nonmalignant_P5_CD3+CD5intSSCint_T4_afterther	3725		√	
		HLA-DR	48		√	
		HLA-DR_control	2397		√	
		CD19	26		√	
		CD19_control	1760		√	
		CD8	5662		√	
		10-empty-cells-in-BC	10		√	
		10-empty-cells-in-DC	10		√	
		10-empty-cells-in-MC	10		√	
		10-empty-cells-in-NK	10		√	
		10-empty-cells-in-TC	10		√	
	BroadS1	Bn	1169	13183	√	
		Bm	491		√	
		DC	142		√	
		M14	1263		√	
		M16	398		√	
		NK	1394		√	
		aTreg	921		√	
		nonT	426		√	
		rTreg	1072		√	
		T4em	975		√	
		T4naive	1134		√	
		T8em	1031		√	
		T8naive	1336		√	
		Tncl	1431		√	
	BroadS2 (Clean)	BC	1884	12292	√	
		DC	202		√	
		pDC	68		√	
		M14	1809		√	
		M16	323		√	
		NK	842		√	
		T4	3380		√	
		T8	3784		√	

Accuracy:	0.059281458			
Precision:	0.68674699	0.03145026	1	0.96851
Recall/Sens	0.005651958	0.99770291	0.000239	0.037286
Specificity:	0.99965489	0.03086546	1	0.9963
F1_Score:	0.01121164	0.06097832	0.000477	0.071807

Predicted	B_cells	Monocytes	NK_cells	T_cells	All
B_cells	57	10028	0	0	10085
Monocytes	1	2606	0	5	2612
NK_cells	20	8290	2	73	8385
T_cells	5	61937	0	2399	64341
All	83	82861	2	2477	85423

True/ Predicted			BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	BC	021-CD19+B___BC	57					10085	0.9943	10085
		021-CD19+B___MC			10028					
Monocytes	M14	003-M14___BC	1					2612	0.0023	2612
		003-M14___MC			2606					
		003-M14___TC					5			
NK_cells	NK	018-CD56+NK___BC	20					8385	0.9998	8385
		018-CD56+NK___MC			8290					
		018-CD56+NK___NK				2				
		018-CD56+NK___TC					73			
T_cells	CD45RA+CD25-T4naive	025-CD4+CD45RA+CD25-NaiveT___BC	1					10479	0.9972	64341
		025-CD4+CD45RA+CD25-NaiveT___MC			10449					
		025-CD4+CD45RA+CD25-NaiveT___TC					29			
	T4	026-T4___BC	1					11213	0.9777	
		026-T4___MC			10962					
		026-T4___TC					250			
	CD45RA+T8naive	027-CD8+CD45RA+NaiveCytotoxicT___MC			11931			11953	0.9982	
		027-CD8+CD45RA+NaiveCytotoxicT___TC					22			
	T8	022-T8___MC			9840			10209	0.9639	
		022-T8___TC					369			
	CD45RO+T4mem	023-CD4+CD45RO+MemoryT___MC			9634			10224	0.9423	
		023-CD4+CD45RO+MemoryT___TC					590			
	CD4+CD25+Treg	024-CD4+CD25+RegulatoryT___BC	3					10263	0.8890	
		024-CD4+CD25+RegulatoryT___MC			9121					
		024-CD4+CD25+RegulatoryT___TC					1139			
All (predicted)			**83**	**0**	**82861**	**2**	**2477**	**85423**		**85423**

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
4	10x (Clean)	BC	10085	85423	V	
		M14	2612		V	
		NK	8385		V	
		CD45RA+CD25-T4naive	10479		V	
		T4	11213		V	
		CD45RA+T8naive	11953		V	
		T8	10269		V	
		CD45RO+T4mem	10224		V	
		CD4+CD25+Treg	10263		V	
	GEO (ALL+10EC*5)	M14_d1	425	34700		V
		M14_d2	431			V
		NK	309			V
		T4	222			V
		T8	310			V
		iNKT	325			V
		MAIT	382			V
		Vd1	284			V
		Vd2	204			V
		T4	965			V
		CCR5+CD69-T4	435			V
		tumor_ascites_DC	1613			V
		tonsil_DC	2739			V
		T8_methanol_SSC	4753			V
		donor1_ IL-10-producing_Foxp3-_T4	1247			V
		donor2_ IL-10-producing_Foxp3-_T4	1902			V
		nonmalignant_P5_CD3+CD5intSSCint_T4	4486			V
		nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy	3725			V
		HLA-DR	48			V
		HLA-DR_control	2397			V
		CD19	26			V
		CD19_control	1760			V
		CD8	5662			V
		10-empty-cells-in-BC	10			V
		10-empty-cells-in-DC	10			V
		10-empty-cells-in-MC	10			V
		10-empty-cells-in-NK	10			V
		10-empty-cells-in-TC	10			V
	BroadS1	Bn	1169	13183	V	
		Bm	491		V	
		DC	142		V	
		M14	1263		V	
		M16	398		V	
		NK	1394		V	
		aTreg	921		V	
		nonT	426		V	
		rTreg	1072		V	
		T4em	975		V	
		T4naive	1134		V	
		T8em	1031		V	
		T8naive	1336		V	
		Tncl	1431		V	
	BroadS2 (Clean)	BC	1884	12292	V	
		DC	202		V	
		pDC	68		V	
		M14	1809		V	
		M16	323		V	
		NK	842		V	
		T4	3380		V	
		T8	3784		V	

Accuracy:	0.751757925				
Precision:	0.6747182	0.19607843	0.36102	0.099451	0.961145
Recall/Sensi	0.69988864	0.00229253	0.731199	0.965517	0.886721
Specificity:	0.98158279	0.99864856	0.863487	0.91888	0.908766
F1_Score:	0.68707297	0.00453206	0.483378	0.180328	0.922434

Predicted	B_cells	ndritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	11	204	79	245	1796
Dendritic_ce	132	10	3748	176	296	4362
Monocytes	64	10	2421	465	351	3311
NK_cells	0	10	0	308	1	319
T_cells	410	10	333	2069	22090	24912
All	1863	51	6706	3097	22983	34700

True/ Predicted			BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	CD19_control	GEO_GSM3258348_CD19_control___BC	1249					1760	0.2903	1796
		GEO_GSM3258348_CD19_control___MC			197					
		GEO_GSM3258348_CD19_control___NK				79				
		GEO_GSM3258348_CD19_control___TC					235			
	CD19	GEO_GSM3258346_CD19___BC	8					26	0.6923	
		GEO_GSM3258346_CD19___DC		1						
		GEO_GSM3258346_CD19___MC			7					
		GEO_GSM3258346_CD19___TC					10			
	10-empty-cells-in-BC	10EC-in-BC___DC		10				10	1.0000	
Dendritic_cells	tonsil_DC	GEO_GSM3162630_tonsil_DC_BC	18					1613	1.0000	4362
		GEO_GSM3162630_tonsil_DC_MC			1420					
		GEO_GSM3162630_tonsil_DC_NK				17				
		GEO_GSM3162630_tonsil_DC_TC					158			
	tumor_ascites_DC	GEO_GSM3162632_tumor_ascites_DC_BC	114					2739	1.0000	
		GEO_GSM3162632_tumor_ascites_DC_MC			2328					
		GEO_GSM3162632_tumor_ascites_DC_NK				159				
		GEO_GSM3162632_tumor_ascites_DC_TC					138			
	10-empty-cells-in-DC	10EC-in-DC___DC		10				10	1.0000	
Monocytes	M14_d1	GEO_GSM2773408_M14_d1___MC			420			425	0.0118	3311
		GEO_GSM2773408_M14_d1___NK				1				
		GEO_GSM2773408_M14_d1___TC					4			
	M14_d2	GEO_GSM2773409_M14_d2___BC	3					431	0.0278	
		GEO_GSM2773409_M14_d2___MC			419					
		GEO_GSM2773409_M14_d2___NK				4				
		GEO_GSM2773409_M14_d2___TC					5			
	HLA-DR	GEO_GSM3258345_HLA-DR_BC	5					48	0.3125	
		GEO_GSM3258345_HLA-DR_MC			33					
		GEO_GSM3258345_HLA-DR_NK				3				
		GEO_GSM3258345_HLA-DR_TC					7			
	HLA-DR_control	GEO_GSM3258347_HLA-DR_control_BC	56					2397	0.3538	
		GEO_GSM3258347_HLA-DR_control_MC			1549					
		GEO_GSM3258347_HLA-DR_control_NK				457				
		GEO_GSM3258347_HLA-DR_control_TC					335			
	10-empty-cells-in-MC	10EC-in-MC___DC		10				10	1.0000	
NK_cells	NK	GEO_GSM3544603_NK___NK				308		309	0.0032	319
		GEO_GSM3544603_NK___TC					1			
	10-empty-cells-in-NK	10EC-in-NK___DC		10				10	1.0000	
T_cells	T4	GEO_20190108_GSM3544603_T4___TC					222	222	0.0000	24912
	T8	GEO_20190108_GSM3544603_T8___MC			1			310	0.0161	
		GEO_20190108_GSM3544603_T8___NK				4				
		GEO_20190108_GSM3544603_T8___TC					305			
	iNKT	GEO_20190108_GSM3544603_iNKT___NK				37		325	0.1138	
		GEO_20190108_GSM3544603_iNKT___TC					288			
	MAIT	GEO_20190108_GSM3544603_MAIT___NK				20		382	0.0524	
		GEO_20190108_GSM3544603_MAIT___TC					362			
	Vd1	GEO_20190108_GSM3544603_Vd1___MC			1			284	0.4542	
		GEO_20190108_GSM3544603_Vd1___NK				128				
		GEO_20190108_GSM3544603_Vd1___TC					155			
	Vd2	GEO_20190108_GSM3544603_Vd2___NK				44		204	0.2157	
		GEO_20190108_GSM3544603_Vd2___TC					160			
	T4	GEO_20190620_GSM3209407_T4___NK				16		965	0.0166	
		GEO_20190620_GSM3209407_T4___TC					949			
	CCR5+CD69-T4	GEO_20190620_GSM3209408_CCR5+CD69-T4___NK				9		435	0.0207	
		GEO_20190620_GSM3209408_CCR5+CD69-T4___TC					426			
	T8_methanol_SSC	GEO_GSM3087629_T8_methanol_SSC_BC	183					4753	0.2981	
		GEO_GSM3087629_T8_methanol_SSC_MC			98					
		GEO_GSM3087629_T8_methanol_SSC_NK				1136				
		GEO_GSM3087629_T8_methanol_SSC_TC					3336			
	donor1_ IL-10-producing_Foxp3-_T4	GEO_GSM3430548_donor1_ IL-10-producing_Foxp3-_T4_NK				6		1247	0.0048	
		GEO_GSM3430548_donor1_ IL-10-producing_Foxp3-_T4_TC					1241			
	donor2_ IL-10-producing_Foxp3-_T4	GEO_GSM3430549_donor2_ IL-10-producing_Foxp3-_T4_BC	1					1902	0.0068	
		GEO_GSM3430549_donor2_ IL-10-producing_Foxp3-_T4_NK				12				
		GEO_GSM3430549_donor2_ IL-10-producing_Foxp3-_T4_TC					1889			
	nonmalignant_P5_CD3+CD5intSSCint_T4	GEO_GSM3478792_nonmalignant_P5_CD3+CD5intSSCint_T4_BC	1					4486	0.0069	
		GEO_GSM3478792_nonmalignant_P5_CD3+CD5intSSCint_T4_MC			22					
		GEO_GSM3478792_nonmalignant_P5_CD3+CD5intSSCint_T4_NK				8				
		GEO_GSM3478792_nonmalignant_P5_CD3+CD5intSSCint_T4_TC					4455			
	nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy	GEO_GSM3558027_nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy_BC	5					3725	0.0030	
		GEO_GSM3558027_nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy_NK				6				
		GEO_GSM3558027_nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy_TC					3714			
	CD8	GEO_GSM3087628_CD8___BC	220					5662	0.1897	
		GEO_GSM3087628_CD8___MC			211					
		GEO_GSM3087628_CD8___NK				643				
		GEO_GSM3087628_CD8___TC					4588			
	10-empty-cells-in-TC	10EC-in-TC___DC		10				10	1.0000	
All (predicted)			**1863**	**51**	**6706**	**3097**	**22983**	**34700**		**34700**

SplitConfusionMatrix-R5

(Compared to R1 (R1 included ALL groups - as the first round), R5 removed the 'EC' group.)

Train: 10x(Clean)+GEO(of R5)+BroadS2(Clean)

Test: BroadS1

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
1	10x (Clean)	BC	10085	85423	√	
		M14	2612		√	
		NK	8385		√	
		CD45RA+CD25-T4naive	10479		√	
		T4	11213		√	
		CD45RA+T8naive	11953		√	
		T8	10209		√	
		CD45RO+T4mem	10224		√	
		CD4+CD25+Treg	10263		√	
	GEO (of R5)	M14_d1	425	34650	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T4	435		√	
		tumor_ascites_DC	1613		√	
		tonsil_DC	2739		√	
		T8_methanol_SSC	4753		√	
		donor1_ IL-10-producing_Foxp3-_T4	1247		√	
		donor2_ IL-10-producing_Foxp3-_T4	1902		√	
		nonmalignant_P5_CD3+CD5intSSCir	4486		√	
		nonmalignant_P5_CD3+CD5intSSCir	3725		√	
		HLA-DR	48		√	
		HLA-DR_control	2397		√	
		CD19	26		√	
		CD19_control	1760		√	
		CD8	5662		√	
	BroadS1	Bn	1169	13183		√
		Bm	491			√
		DC	142			√
		M14	1263			√
		M16	398			√
		NK	1394			√
		aTreg	921			√
		nonT	426			√
		rTreg	1072			√
		T4em	975			√
		T4naive	1134			√
		T8em	1031			√
		T8naive	1336			√
		Tncl	1431			√
	BroadS2 (Clean)	BC	1884	12292	√	
		DC	202		√	
		pDC	68		√	
		M14	1809		√	
		M16	323		√	
		NK	842		√	
		T4	3380		√	
		T8	3784		√	

Accuracy:	0.936888417					
Precision:	0.99607843	0.80152672	0.94709302	0.79646697	0.9477647	
Recall/Sensitivi	0.91807229	0.73943662	0.9807345	0.74390244	0.9675715	
Specificity:	0.9994793	0.99800629	0.99210207	0.97752142	0.9085856	
F1_Score:	0.95548589	0.76923077	0.96362023	0.76928783	0.9575657	
Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1524	16	42	2	76	1660
Dendritic_cells	0	105	35	0	2	142
Monocytes	4	9	1629	0	19	1661
NK_cells	2	0	8	1037	347	1394
T_cells	0	1	6	263	8056	8326
All	1530	131	1720	1302	8500	13183

True/ Predicted					BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	Bn	Bn_aTreg	BT580	Bn_aTreg_BT580___BC	4					1169	0.0796	1660
			BT860	Bn_aTreg_BT860___BC	6							
			NY860	Bn_aTreg_NY860___BC	3							
		Bn_nonT	BT580	Bn_nonT_BT580___BC	233							
				Bn_nonT_BT580___DC		1						
				Bn_nonT_BT580___MC			8					
				Bn_nonT_BT580___TC					5			
			BT860	Bn_nonT_BT860___BC	512							
				Bn_nonT_BT860___DC		4						
				Bn_nonT_BT860___MC			17					
				Bn_nonT_BT860___TC					20			
			NY580	Bn_nonT_NY580___BC	150							
				Bn_nonT_NY580___DC		2						
				Bn_nonT_NY580___MC			1					
				Bn_nonT_NY580___TC					11			
			NY860	Bn_nonT_NY860___BC	166							
				Bn_nonT_NY860___DC		3						
				Bn_nonT_NY860___MC			5					
				Bn_nonT_NY860___NK				1				
				Bn_nonT_NY860___TC					15			
		Bn_T4em	BT860	Bn_T4em_BT860___BC	1							
		Bn_Tncl	BT860	Bn_Tncl_BT860___BC	1							
	Bm	Bm_aTreg	BT860	Bm_aTreg_BT860___BC	6					491	0.0876	
			NY580	Bm_aTreg_NY580___BC	1							
			NY860	Bm_aTreg_NY860___BC	2							
		Bm_nonT	BT580	Bm_nonT_BT580___BC	85							
				Bm_nonT_BT580___MC			3					
			BT860	Bm_nonT_BT860___BC	208							
				Bm_nonT_BT860___DC		2						
				Bm_nonT_BT860___MC			4					
				Bm_nonT_BT860___TC					9			
			NY580	Bm_nonT_NY580___BC	59							
				Bm_nonT_NY580___DC		1						
				Bm_nonT_NY580___TC					3			
			NY860	Bm_nonT_NY860___BC	87							
				Bm_nonT_NY860___DC		3						
				Bm_nonT_NY860___MC			4					
				Bm_nonT_NY860___NK				1				
				Bm_nonT_NY860___TC					13			
Dendritic_cells	DC	DC_aTreg	BT860	DC_aTreg_BT860___DC		1				142	0.2606	142
			NY580	DC_aTreg_NY580___DC		1						
		DC_nonT	BT580	DC_nonT_BT580___DC		36						
				DC_nonT_BT580___MC			18					
			BT860	DC_nonT_BT860___DC		14						
				DC_nonT_BT860___MC			4					
				DC_nonT_BT860___TC					1			
			NY580	DC_nonT_NY580___DC		38						
				DC_nonT_NY580___MC			8					
			NY860	DC_nonT_NY860___DC		15						
				DC_nonT_NY860___MC			5					
				DC_nonT_NY860___TC					1			
Monocytes	M14	M14_aTreg	BT580	M14_aTreg_BT580___MC			1			1263	0.0182	1661
			BT860	M14_aTreg_BT860___MC			4					
			NY580	M14_aTreg_NY580___MC			2					
			NY860	M14_aTreg_NY860___MC			2					
		M14_nonT	BT580	M14_nonT_BT580___BC	1							
				M14_nonT_BT580___DC		1						
				M14_nonT_BT580___MC			234					
				M14_nonT_BT580___TC					2			
			BT860	M14_nonT_BT860___BC	2							
				M14_nonT_BT860___DC		4						
				M14_nonT_BT860___MC			328					
				M14_nonT_BT860___TC					4			
			NY580	M14_nonT_NY580___MC			339					
				M14_nonT_NY580___TC					2			
			NY860	M14_nonT_NY860___MC			328					
				M14_nonT_NY860___TC					7			
		M14_rTreg	NY580	M14_rTreg_NY580___MC			1					
		M14_Tncl	BT580	M14_Tncl_BT580___MC			1					
	M16	M16_aTreg	BT580	M16_aTreg_BT580___MC			4			398	0.0226	
			BT860	M16_aTreg_BT860___MC			5					
			NY580	M16_aTreg_NY580___MC			7					
			NY860	M16_aTreg_NY860___MC			7					
		M16_nonT	BT580	M16_nonT_BT580___DC		2						
				M16_nonT_BT580___MC			57					
			BT860	M16_nonT_BT860___BC	1							
				M16_nonT_BT860___DC		1						
				M16_nonT_BT860___MC			101					
				M16_nonT_BT860___TC					4			
			NY580	M16_nonT_NY580___MC			81					
			NY860	M16_nonT_NY860___DC		1						
				M16_nonT_NY860___MC			125					
		M16_T8em	BT580	M16_T8em_BT580___MC			1					
		M16_T8em	NY860	M16_T8em_NY860___MC			1					

NK_cells	NK	NK_aTreg	BT580	NK_aTreg_BT580__TC					2	1394	0.2561	1394
			NY580	NK_aTreg_NY580__TC					3			
			NY860	NK_aTreg_NY860__TC					1			
		NK_nonT	BT580	NK_nonT_BT580__MC			3					
				NK_nonT_BT580__NK				225				
				NK_nonT_BT580__TC					26			
			BT860	NK_nonT_BT860__BC	2							
				NK_nonT_BT860__MC			1					
				NK_nonT_BT860__NK				340				
				NK_nonT_BT860__TC					82			
			NY580	NK_nonT_NY580__MC			2					
				NK_nonT_NY580__NK				157				
				NK_nonT_NY580__TC					29			
			NY860	NK_nonT_NY860__NK				219				
				NK_nonT_NY860__TC					45			
		NK_T4em	NY860	NK_T4em_NY860__TC					1			
		NK_T4naive	NY860	NK_T4naive_NY860__TC					1			
		NK_T8em	BT580	NK_T8em_BT580__MC			1					
				NK_T8em_BT580__NK				13				
				NK_T8em_BT580__TC					26			
			BT860	NK_T8em_BT860__MC			1					
				NK_T8em_BT860__NK				28				
				NK_T8em_BT860__TC					57			
			NY580	NK_T8em_NY580__NK				11				
				NK_T8em_NY580__TC					7			
			NY860	NK_T8em_NY860__NK				33				
				NK_T8em_NY860__TC					35			
		NK_Tncl	BT580	NK_Tncl_BT580__NK				3				
				NK_Tncl_BT580__TC					7			
			BT860	NK_Tncl_BT860__NK				4				
				NK_Tncl_BT860__TC					6			
			NY580	NK_Tncl_NY580__NK				1				
				NK_Tncl_NY580__TC					10			
			NY860	NK_Tncl_NY860__NK				3				
				NK_Tncl_NY860__TC					9			
T_cells	aTreg	T_aTreg	BT580	T_aTreg_BT580__NK				1		921	0.0011	8326
				T_aTreg_BT580__TC					240			
			BT860	T_aTreg_BT860__TC					243			
			NY580	T_aTreg_NY580__TC					222			
			NY860	T_aTreg_NY860__TC					215			
	nonT	T_nonT	BT580	T_nonT_BT580__MC			1			426	0.4484	
				T_nonT_BT580__NK				45				
				T_nonT_BT580__TC					50			
			BT860	T_nonT_BT860__MC			1					
				T_nonT_BT860__NK				51				
				T_nonT_BT860__TC					82			
			NY580	T_nonT_NY580__NK				48				
				T_nonT_NY580__TC					37			
			NY860	T_nonT_NY860__NK				45				
				T_nonT_NY860__TC					66			
	rTreg	T_rTreg	BT580	T_rTreg_BT580__MC			1			1072	0.0047	
				T_rTreg_BT580__NK				1				
				T_rTreg_BT580__TC					311			
			BT860	T_rTreg_BT860__MC			1					
				T_rTreg_BT860__NK				1				
				T_rTreg_BT860__TC					232			
			NY580	T_rTreg_NY580__TC					337			
			NY860	T_rTreg_NY860__NK				1				
				T_rTreg_NY860__TC					187			
	T4em	T_T4em	BT580	T_T4em_BT580__MC			1			975	0.0072	
				T_T4em_BT580__NK				4				
				T_T4em_BT580__TC					325			
			BT860	T_T4em_BT860__NK				1				
				T_T4em_BT860__TC					258			
			NY580	T_T4em_NY580__NK				1				
				T_T4em_NY580__TC					253			
			NY860	T_T4em_NY860__TC					132			
	T4naive	T_T4naive	BT580	T_T4naive_BT580__DC		1				1134	0.0018	
				T_T4naive_BT580__NK				1				
				T_T4naive_BT580__TC					480			
			BT860	T_T4naive_BT860__TC					265			
			NY580	T_T4naive_NY580__TC					291			
			NY860	T_T4naive_NY860__TC					96			
	T8em	T_T8em	BT580	T_T8em_BT580__NK				12		1031	0.0398	
				T_T8em_BT580__TC					254			
			BT860	T_T8em_BT860__NK				13				
				T_T8em_BT860__TC					291			
			NY580	T_T8em_NY580__NK				9				
				T_T8em_NY580__TC					257			
			NY860	T_T8em_NY860__NK				7				
				T_T8em_NY860__TC					188			
	T8naive	T_T8naive	BT580	T_T8naive_BT580__TC					318	1336	0.0000	
			BT860	T_T8naive_BT860__TC					486			
			NY580	T_T8naive_NY580__TC					256			
			NY860	T_T8naive_NY860__TC					276			
	Tncl	T_Tncl	BT580	T_Tncl_BT580__MC			1			1431	0.0161	
				T_Tncl_BT580__NK				5				
				T_Tncl_BT580__TC					195			
			BT860	T_Tncl_BT860__NK				3				
				T_Tncl_BT860__TC					363			
			NY580	T_Tncl_NY580__NK				7				
				T_Tncl_NY580__TC					372			
			NY860	T_Tncl_NY860__NK				7				
				T_Tncl_NY860__TC					478			
All (predicted)					**1530**	**131**	**1720**	**1302**	**8500**	**13183**		**13183**

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
2	10x (Clean)	BC	10085	85423	√	
		M14	2612		√	
		NK	8385		√	
		CD45RA+CD2	10479		√	
		T4	11213		√	
		CD45RA+T8n	11953		√	
		T8	10209		√	
		CD45RO+T4m	10224		√	
		CD4+CD25+T	10263		√	
	GEO (of R5)	M14_d1	425	34650	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T	435		√	
		tumor_ascite	1613		√	
		tonsil_DC	2739		√	
		T8_methano	4753		√	
		donor1_ IL-1	1247		√	
		donor2_ IL-1	1902		√	
		nonmalignan	4486		√	
		nonmalignan	3725		√	
		HLA-DR	48		√	
		HLA-DR_cont	2397		√	
		CD19	26		√	
		CD19_contro	1760		√	
		CD8	5662		√	
	BroadS1	Bn	1169	13183	√	
		Bm	491		√	
		DC	142		√	
		M14	1263		√	
		M16	398		√	
		NK	1394		√	
		aTreg	921		√	
		nonT	426		√	
		rTreg	1072		√	
		T4em	975		√	
		T4naive	1134		√	
		T8em	1031		√	
		T8naive	1336		√	
		Tncl	1431		√	
	BroadS2 (Clean)	BC	1884	12292		√
		DC	202			√
		pDC	68			√
		M14	1809			√
		M16	323			√
		NK	842			√
		T4	3380			√
		T8	3784			√

Accuracy:	0.910023				
Precision:	0.94925	0.714286	0.8980322	0.67121849	0.938041
Recall/Ser	0.873673	0.240741	0.94183865	0.75890736	0.953099
Specificity	0.991545	0.997837	0.97755906	0.97266376	0.912051
F1_Score:	0.909895	0.360111	0.91941392	0.71237458	0.94551

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1646	10	45	0	183	1884
Dendritic_	17	65	166	4	18	270
Monocyte	62	12	2008	0	50	2132
NK_cells	0	1	2	639	200	842
T_cells	9	3	15	309	6828	7164
All	1734	91	2236	952	7279	12292

True/ Predicted						BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	BC	pbmc1	v2	A	pbmc1_v2_A_BC___BC	227					1884	0.1263	1884
					pbmc1_v2_A_BC___DC		4						
					pbmc1_v2_A_BC___MC			11					
					pbmc1_v2_A_BC___TC					46			
				B	pbmc1_v2_B_BC___BC	287							
					pbmc1_v2_B_BC___MC			15					
					pbmc1_v2_B_BC___TC					86			
			v3		pbmc1_v3_BC___BC	303							
					pbmc1_v3_BC___MC			14					
					pbmc1_v3_BC___TC					29			
		pbmc2	v2		pbmc2_V2_BC___BC	829							
					pbmc2_V2_BC___DC		6						
					pbmc2_V2_BC___MC			5					
					pbmc2_V2_BC___TC					22			
Dendritic_cells	DC	pbmc1	v2	A	pbmc1_v2_A_DC___BC	1					202	0.6980	270
					pbmc1_v2_A_DC___DC		11						
					pbmc1_v2_A_DC___MC			41					
					pbmc1_v2_A_DC___TC					2			
				B	pbmc1_v2_B_DC___DC		1						
					pbmc1_v2_B_DC___MC			31					
					pbmc1_v2_B_DC___TC					1			
			v3		pbmc1_v3_DC___BC	1							
					pbmc1_v3_DC___DC		1						
					pbmc1_v3_DC___MC			32					
					pbmc1_v3_DC___NK				1				
					pbmc1_v3_DC___TC					3			
		pbmc2	v2		pbmc2_V2_DC___BC	2							
					pbmc2_V2_DC___DC		48						
					pbmc2_V2_DC___MC			23					
					pbmc2_V2_DC___TC					3			
	pDC	pbmc1	v2	A	pbmc1_v2_A_pDC___BC	7					68	0.9412	
					pbmc1_v2_A_pDC___MC			13					
					pbmc1_v2_A_pDC___NK				1				
					pbmc1_v2_A_pDC___TC					5			
				B	pbmc1_v2_B_pDC___MC			9					
					pbmc1_v2_B_pDC___TC					3			
		pbmc2	V2		pbmc2_V2_pDC___BC	6							
					pbmc2_V2_pDC___DC		4						
					pbmc2_V2_pDC___MC			17					
					pbmc2_V2_pDC___NK				2				
					pbmc2_V2_pDC___TC					1			
Monocytes	M14	pbmc1	v2	A	pbmc1_v2_A_M14___BC	22					1809	0.0641	2132
					pbmc1_v2_A_M14___DC		6						
					pbmc1_v2_A_M14___MC			601					
					pbmc1_v2_A_M14___TC					11			
				B	pbmc1_v2_B_M14___BC	2							
					pbmc1_v2_B_M14___MC			372					
					pbmc1_v2_B_M14___TC					5			
			v3		pbmc1_v3_M14___BC	5							
					pbmc1_v3_M14___MC			340					
					pbmc1_v3_M14___TC					9			
		pbmc2	V2		pbmc2_V2_M14___BC	31							
					pbmc2_V2_M14___DC		5						
					pbmc2_V2_M14___MC			380					
					pbmc2_V2_M14___TC					20			
	M16	pbmc1	v2	A	pbmc1_v2_A_M16___BC	1					323	0.0248	
					pbmc1_v2_A_M16___DC		1						
					pbmc1_v2_A_M16___MC			96					
					pbmc1_v2_A_M16___TC					4			
				B	pbmc1_v2_B_M16___MC			73					
			v3		pbmc1_v3_M16___BC	1							
					pbmc1_v3_M16___MC			96					
					pbmc1_v3_M16___TC					1			
		pbmc2	V2		pbmc2_V2_M16___MC			50					
NK_cells	NK	pbmc1	v2	A	pbmc1_v2_A_NK___MC			1			842	0.2411	842
					pbmc1_v2_A_NK___NK				131				
					pbmc1_v2_A_NK___TC					34			
				B	pbmc1_v2_B_NK___MC			1					
					pbmc1_v2_B_NK___NK				169				
					pbmc1_v2_B_NK___TC					93			
			v3		pbmc1_v3_NK___NK				130				
					pbmc1_v3_NK___TC					64			
		pbmc2	V2		pbmc2_V2_NK___DC		1						
					pbmc2_V2_NK___NK				209				
					pbmc2_V2_NK___TC					9			
T_cells	T4	pbmc1	v2	A	pbmc1_v2_A_T4___BC	2					3380	0.0133	7164
					pbmc1_v2_A_T4___DC		1						
					pbmc1_v2_A_T4___NK				8				
					pbmc1_v2_A_T4___TC					539			
				B	pbmc1_v2_B_T4___MC			4					
					pbmc1_v2_B_T4___NK				6				
					pbmc1_v2_B_T4___TC					898			
			v3		pbmc1_v3_T4___NK				14				
					pbmc1_v3_T4___TC					946			
		pbmc2	V2		pbmc2_V2_T4___BC	1							
					pbmc2_V2_T4___DC		2						
					pbmc2_V2_T4___MC			6					
					pbmc2_V2_T4___NK				1				
					pbmc2_V2_T4___TC					952			
	T8	pbmc1	v2	A	pbmc1_v2_A_T8___BC	3					3784	0.0769	
					pbmc1_v2_A_T8___MC			3					
					pbmc1_v2_A_T8___NK				114				
					pbmc1_v2_A_T8___TC					1054			
				B	pbmc1_v2_B_T8___MC			2					
					pbmc1_v2_B_T8___NK				57				
					pbmc1_v2_B_T8___TC					895			
			v3		pbmc1_v3_T8___NK				28				
					pbmc1_v3_T8___TC					934			
		pbmc2	V2		pbmc2_V2_T8___BC	3							
					pbmc2_V2_T8___NK				81				
					pbmc2_V2_T8___TC					610			
All (predicted)						1734	91	2236	952	7279	12292		12292

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
3	10x (Clean)	BC	10085	85423		√
		M14	2612			√
		NK	8385			√
		CD45RA+CD25-T4naive	10479			√
		T4	11213			√
		CD45RA+T8naive	11953			√
		T8	10209			√
		CD45RO+T4mem	10224			√
		CD4+CD25+Treg	10263			√
	GEO (of R5)	M14_d1	425	34650	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T4	435		√	
		tumor_ascites_DC	1613		√	
		tonsil_DC	2739		√	
		T8_methanol_SSC	4753		√	
		donor1_ IL-10-producing_Foxp3-_T4	1247		√	
		donor2_ IL-10-producing_Foxp3-_T4	1902		√	
		nonmalignant_P5_CD3+CD5intSSCint_T4	4486		√	
		nonmalignant_P5_CD3+CD5intSSCint_T4_afterther	3725		√	
		HLA-DR	48		√	
		HLA-DR_control	2397		√	
		CD19	26		√	
		CD19_control	1760		√	
		CD8	5662		√	
	BroadS1	Bn	1169	13183	√	
		Bm	491		√	
		DC	142		√	
		M14	1263		√	
		M16	398		√	
		NK	1394		√	
		aTreg	921		√	
		nonT	426		√	
		rTreg	1072		√	
		T4em	975		√	
		T4naive	1134		√	
		T8em	1031		√	
		T8naive	1336		√	
		Tncl	1431		√	
	BroadS2 (Clean)	BC	1884	12292	√	
		DC	202		√	
		pDC	68		√	
		M14	1809		√	
		M16	323		√	
		NK	842		√	
		T4	3380		√	
		T8	3784		√	

Accuracy:	0.128162205				
Precision:	0.49921916	0	0.031808	1	0.945082
Recall/Sens	0.09509172	0	0.918836	0.030769	0.11394
Specificity:	0.98723088	0.99959027	0.117847	1	0.979793
F1_Score:	0.15975346	0	0.061488	0.059701	0.203362

Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	959	0	9121	0	5	10085
Monocytes	3	34	2400	0	175	2612
NK_cells	6	0	7875	258	246	8385
T_cells	953	1	56056	0	7331	64341
All	1921	35	75452	258	7757	85423

True/ Predicted			BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	BC	021-CD19+B___BC	959					10085	0.9049	10085
		021-CD19+B___MC			9121					
		021-CD19+B___TC					5			
Monocytes	M14	003-M14___BC	3					2612	0.0812	2612
		003-M14___DC		34						
		003-M14___MC			2400					
		003-M14___TC					175			
NK_cells	NK	018-CD56+NK___BC	6					8385	0.9692	8385
		018-CD56+NK___MC			7875					
		018-CD56+NK___NK				258				
		018-CD56+NK___TC					246			
T_cells	CD45RA+CD25-T4naive	025-CD4+CD45RA+CD25-NaiveT___BC	270					10479	0.9584	64341
		025-CD4+CD45RA+CD25-NaiveT___MC			9773					
		025-CD4+CD45RA+CD25-NaiveT___TC					436			
	T4	026-T4___BC	241					11213	0.9452	
		026-T4___MC			10358					
		026-T4___TC					614			
	CD45RA+T8naive	027-CD8+CD45RA+NaiveCytotoxicT___BC	9					11953	0.9483	
		027-CD8+CD45RA+NaiveCytotoxicT___MC			11326					
		027-CD8+CD45RA+NaiveCytotoxicT___TC					618			
	T8	022-T8___BC	8					10209	0.8027	
		022-T8___MC			8187					
		022-T8___TC					2014			
	CD45RO+T4mem	023-CD4+CD45RO+MemoryT___BC	18					10224	0.8353	
		023-CD4+CD45RO+MemoryT___DC		1						
		023-CD4+CD45RO+MemoryT___MC			8521					
		023-CD4+CD45RO+MemoryT___TC					1684			
	CD4+CD25+Treg	024-CD4+CD25+RegulatoryT___BC	407					10263	0.8085	
		024-CD4+CD25+RegulatoryT___MC			7891					
		024-CD4+CD25+RegulatoryT___TC					1965			
All (predicted)			**1921**	**35**	**75452**	**258**	**7757**	**85423**		**85423**

EXP	DataSets	Subtype	SubtypeN	TotalCell	Training	Testing
4	10x (Clean)	BC	10085	85423	√	
		M14	2612		√	
		NK	8385		√	
		CD45RA+CD25-T4naive	10479		√	
		T4	11213		√	
		CD45RA+T8naive	11953		√	
		T8	10209		√	
		CD45RO+T4mem	10224		√	
		CD4+CD25+Treg	10263		√	
	GEO (of R5)	M14_d1	425	34650		√
		M14_d2	431			√
		NK	309			√
		T4	222			√
		T8	310			√
		iNKT	325			√
		MAIT	382			√
		Vd1	284			√
		Vd2	204			√
		T4	965			√
		CCR5+CD69-T4	435			√
		tumor_ascites_DC	1613			√
		tonsil_DC	2739			√
		T8_methanol_SSC	4753			√
		donor1_ IL-10-producing_Foxp3-_T4	1247			√
		donor2_ IL-10-producing_Foxp3-_T4	1902			√
		nonmalignant_P5_CD3+CD5intSSCint_T4	4486			√
		nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy	3725			√
		HLA-DR	48			√
		HLA-DR_control	2397			√
		CD19	26			√
		CD19_control	1760			√
		CD8	5662			√
	BroadS1	Bn	1169	13183	√	
		Bm	491		√	
		DC	142		√	
		M14	1263		√	
		M16	398		√	
		NK	1394		√	
		aTreg	921		√	
		nonT	426		√	
		rTreg	1072		√	
		T4em	975		√	
		T4naive	1134		√	
		T8em	1031		√	
		T8naive	1336		√	
		Tncl	1431		√	
	BroadS2 (Clean)	BC	1884	12292	√	
		DC	202		√	
		pDC	68		√	
		M14	1809		√	
		M16	323		√	
		NK	842		√	
		T4	3380		√	
		T8	3784		√	

Accuracy:	0.75255411				
Precision:	0.6747182	0	0.36102	0.099451	0.961145
Recall/Sensi	0.70380739	0	0.733414	0.996764	0.887077
Specificity:	0.98156037	0.999967	0.863313	0.918785	0.908391
F1_Score:	0.68895588	0	0.483861	0.180857	0.922627

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	1	204	79	245	1786
Dendritic_ce	132	0	3748	176	296	4352
Monocytes	64	0	2421	465	351	3301
NK_cells	0	0	0	308	1	309
T_cells	410	0	333	2069	22090	24902
All	1863	1	6706	3097	22983	34650

True/ Predicted			BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	CD19_control	GEO_GSM3258348_CD19_control___BC	1249					1760	0.2903	1786
		GEO_GSM3258348_CD19_control___MC			197					
		GEO_GSM3258348_CD19_control___NK				79				
		GEO_GSM3258348_CD19_control___TC					235			
	CD19	GEO_GSM3258346_CD19___BC	8					26	0.6923	
		GEO_GSM3258346_CD19___DC		1						
		GEO_GSM3258346_CD19___MC			7					
		GEO_GSM3258346_CD19___TC					10			
Dendritic_cells	tonsil_DC	GEO_GSM3162630_tonsil_DC_BC	18					1613	1.0000	4352
		GEO_GSM3162630_tonsil_DC_MC			1420					
		GEO_GSM3162630_tonsil_DC_NK				17				
		GEO_GSM3162630_tonsil_DC_TC					158			
	tumor_ascites_DC	GEO_GSM3162632_tumor_ascites_DC_BC	114					2739	1.0000	
		GEO_GSM3162632_tumor_ascites_DC_MC			2328					
		GEO_GSM3162632_tumor_ascites_DC_NK				159				
		GEO_GSM3162632_tumor_ascites_DC_TC					138			
Monocytes	M14_d1	GEO_GSM2773408_M14_d1___MC			420			425	0.0118	3301
		GEO_GSM2773408_M14_d1___NK				1				
		GEO_GSM2773408_M14_d1___TC					4			
	M14_d2	GEO_GSM2773409_M14_d2___BC	3					431	0.0278	
		GEO_GSM2773409_M14_d2___MC			419					
		GEO_GSM2773409_M14_d2___NK				4				
		GEO_GSM2773409_M14_d2___TC					5			
	HLA-DR	GEO_GSM3258345_HLA-DR_BC	5					48	0.3125	
		GEO_GSM3258345_HLA-DR_MC			33					
		GEO_GSM3258345_HLA-DR_NK				3				
		GEO_GSM3258345_HLA-DR_TC					7			
	HLA-DR_control	GEO_GSM3258347_HLA-DR_control_BC	56					2397	0.3538	
		GEO_GSM3258347_HLA-DR_control_MC			1549					
		GEO_GSM3258347_HLA-DR_control_NK				457				
		GEO_GSM3258347_HLA-DR_control_TC					335			
NK_cells	NK	GEO_GSM3544603_NK___NK				308		309	0.0032	309
		GEO_GSM3544603_NK___TC					1			
T_cells	T4	GEO_20190108_GSM3544603_T4___TC					222	222	0.0000	24902
	T8	GEO_20190108_GSM3544603_T8___MC			1			310	0.0161	
		GEO_20190108_GSM3544603_T8___NK				4				
		GEO_20190108_GSM3544603_T8___TC					305			
	iNKT	GEO_20190108_GSM3544603_iNKT___NK				37		325	0.1138	
		GEO_20190108_GSM3544603_iNKT___TC					288			
	MAIT	GEO_20190108_GSM3544603_MAIT___NK				20		382	0.0524	
		GEO_20190108_GSM3544603_MAIT___TC					362			
	Vd1	GEO_20190108_GSM3544603_Vd1___MC			1			284	0.4542	
		GEO_20190108_GSM3544603_Vd1___NK				128				
		GEO_20190108_GSM3544603_Vd1___TC					155			
	Vd2	GEO_20190108_GSM3544603_Vd2___NK				44		204	0.2157	
		GEO_20190108_GSM3544603_Vd2___TC					160			
	T4	GEO_20190620_GSM3209407_T4___NK				16		965	0.0166	
		GEO_20190620_GSM3209407_T4___TC					949			
	CCR5+CD69-T4	GEO_20190620_GSM3209408_CCR5+CD69-T4___NK				9		435	0.0207	
		GEO_20190620_GSM3209408_CCR5+CD69-T4___TC					426			
	T8_methanol_SSC	GEO_GSM3087629_T8_methanol_SSC_BC	183					4753	0.2981	
		GEO_GSM3087629_T8_methanol_SSC_MC			98					
		GEO_GSM3087629_T8_methanol_SSC_NK				1136				
		GEO_GSM3087629_T8_methanol_SSC_TC					3336			
	onor1_ IL-10-producing_Foxp3-_T	GEO_GSM3430548_donor1_ IL-10-producing_Foxp3-_T4_NK				6		1247	0.0048	
		GEO_GSM3430548_donor1_ IL-10-producing_Foxp3-_T4_TC					1241			
	onor2_ IL-10-producing_Foxp3-_T	GEO_GSM3430549_donor2_ IL-10-producing_Foxp3-_T4_BC	1					1902	0.0068	
		GEO_GSM3430549_donor2_ IL-10-producing_Foxp3-_T4_NK				12				
		GEO_GSM3430549_donor2_ IL-10-producing_Foxp3-_T4_TC					1889			
	nmalignant_P5_CD3+CD5intSSCint	GEO_GSM3478792_nonmalignant_P5_CD3+CD5intSSCint_T4_BC	1					4486	0.0069	
		GEO_GSM3478792_nonmalignant_P5_CD3+CD5intSSCint_T4_MC			22					
		GEO_GSM3478792_nonmalignant_P5_CD3+CD5intSSCint_T4_NK				8				
		GEO_GSM3478792_nonmalignant_P5_CD3+CD5intSSCint_T4_TC					4455			
	nant_P5_CD3+CD5intSSCint_T4_aft	GEO_GSM3558027_nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy_BC	5					3725	0.0030	
		GEO_GSM3558027_nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy_NK				6				
		GEO_GSM3558027_nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy_TC					3714			
	CD8	GEO_GSM3087628_T8___BC	220					5662	0.1897	
		GEO_GSM3087628_T8___MC			211					
		GEO_GSM3087628_T8___NK				643				
		GEO_GSM3087628_T8___TC					4588			
All (predicted)			**1863**	**1**	**6706**	**3097**	**22983**	**34650**	**5.0906**	**34650**

SplitConfusionMatrix-R7

(Compared to R1 (R1 included ALL groups), R7 removed the 'EC' group and the 'Other Tissue' group.)

Train: 10x(Clean)+GEO(of R7)+BroadS2(Clean)

Test: BroadS1

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
1	10x (Clean)	BC	10085	85423	√	
		M14	2612		√	
		NK	8385		√	
		CD45RA+CD25-T4naive	10479		√	
		T4	11213		√	
		CD45RA+T8naive	11953		√	
		T8	10209		√	
		CD45RO+T4mem	10224		√	
		CD4+CD25+Treg	10263		√	
	GEO (of R7)	M14_d1	425	30298	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T4	435		√	
					√	
					√	
		T8_methanol_SSC	4753		√	
		donor1_ IL-10-producing_Foxp3-_T4	1247		√	
		donor2_ IL-10-producing_Foxp3-_T4	1902		√	
		nonmalignant_P5_CD3+CD5intSSCir	4486		√	
		nonmalignant_P5_CD3+CD5intSSCir	3725		√	
		HLA-DR	48		√	
		HLA-DR_control	2397		√	
		CD19	26		√	
		CD19_control	1760		√	
		CD8	5662		√	
	BroadS1	Bn	1169	13183		√
		Bm	491			√
		DC	142			√
		M14	1263			√
		M16	398			√
		NK	1394			√
		aTreg	921			√
		nonT	426			√
		rTreg	1072			√
		T4em	975			√
		T4naive	1134			√
		T8em	1031			√
		T8naive	1336			√
		Tncl	1431			√
	BroadS2 (Clean)	BC	1884	12292	√	
		DC	202		√	
		pDC	68		√	
		M14	1809		√	
		M16	323		√	
		NK	842		√	
		T4	3380		√	
		T8	3784		√	

Accuracy:	0.941136312					
Precision:	0.99737015	0.83870968	0.9702381	0.74374177	0.9629318	
Recall/Sensitivi	0.91385542	0.91549296	0.98133654	0.80989957	0.9609657	
Specificity:	0.99965287	0.99808297	0.99566048	0.96700314	0.9365864	
F1_Score:	0.95378812	0.87542088	0.97575576	0.77541209	0.9619477	
Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1517	14	37	67	25	1660
Dendritic_cells	0	130	10	0	2	142
Monocytes	1	10	1630	0	20	1661
NK_cells	2	0	2	1129	261	1394
T_cells	1	1	1	322	8001	8326
All	1521	155	1680	1518	8309	13183

True/ Predicted					BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	Bn	Bn_aTreg	BT580	Bn_aTreg_BT580___BC	4					1169	0.0838	1660
			BT860	Bn_aTreg_BT860___BC	6							
			NY860	Bn_aTreg_NY860___BC	2							
				Bn_aTreg_NY860___MC			1					
		Bn_nonT	BT580	Bn_nonT_BT580___BC	237							
				Bn_nonT_BT580___DC		1						
				Bn_nonT_BT580___MC			2					
				Bn_nonT_BT580___NK				5				
				Bn_nonT_BT580___TC					2			
			BT860	Bn_nonT_BT860___BC	509							
				Bn_nonT_BT860___DC		4						
				Bn_nonT_BT860___MC			17					
				Bn_nonT_BT860___NK				19				
				Bn_nonT_BT860___TC					4			
			NY580	Bn_nonT_NY580___BC	147							
				Bn_nonT_NY580___DC		2						
				Bn_nonT_NY580___MC			4					
				Bn_nonT_NY580___NK				6				
				Bn_nonT_NY580___TC					5			
			NY860	Bn_nonT_NY860___BC	164							
				Bn_nonT_NY860___DC		2						
				Bn_nonT_NY860___MC			3					
				Bn_nonT_NY860___NK				16				
				Bn_nonT_NY860___TC					5			
		Bn_T4em	BT860	Bn_T4em_BT860___BC	1							
		Bn_Tncl	BT860	Bn_Tncl_BT860___BC	1							
	Bm	Bm_aTreg	BT860	Bm_aTreg_BT860___BC	6					491	0.0916	
			NY580	Bm_aTreg_NY580___BC	1							
			NY860	Bm_aTreg_NY860___BC	2							
		Bm_nonT	BT580	Bm_nonT_BT580___BC	87							
				Bm_nonT_BT580___MC			1					
			BT860	Bm_nonT_BT860___BC	205							
				Bm_nonT_BT860___DC		2						
				Bm_nonT_BT860___MC			6					
				Bm_nonT_BT860___NK				5				
				Bm_nonT_BT860___TC					5			
			NY580	Bm_nonT_NY580___BC	59							
				Bm_nonT_NY580___DC		1						
				Bm_nonT_NY580___NK				1				
				Bm_nonT_NY580___TC					2			
			NY860	Bm_nonT_NY860___BC	86							
				Bm_nonT_NY860___DC		2						
				Bm_nonT_NY860___MC			3					
				Bm_nonT_NY860___NK				15				
				Bm_nonT_NY860___TC					2			
Dendritic_cells	DC	DC_aTreg	BT860	DC_aTreg_BT860___DC		1				142	0.0845	142
			NY580	DC_aTreg_NY580___DC		1						
		DC_nonT	BT580	DC_nonT_BT580___DC		51						
				DC_nonT_BT580___MC			3					
			BT860	DC_nonT_BT860___DC		16						
				DC_nonT_BT860___MC			2					
				DC_nonT_BT860___TC					1			
			NY580	DC_nonT_NY580___DC		44						
				DC_nonT_NY580___MC			2					
			NY860	DC_nonT_NY860___DC		17						
				DC_nonT_NY860___MC			3					
				DC_nonT_NY860___TC					1			
Monocytes	M14	M14_aTreg	BT580	M14_aTreg_BT580___MC			1			1263	0.0158	1661
			BT860	M14_aTreg_BT860___MC			4					
			NY580	M14_aTreg_NY580___MC			2					
			NY860	M14_aTreg_NY860___MC			2					
		M14_nonT	BT580	M14_nonT_BT580___DC		1						
				M14_nonT_BT580___MC			235					
				M14_nonT_BT580___TC					2			
			BT860	M14_nonT_BT860___BC	1							
				M14_nonT_BT860___DC		4						
				M14_nonT_BT860___MC			328					
				M14_nonT_BT860___TC					5			
			NY580	M14_nonT_NY580___MC			339					
				M14_nonT_NY580___TC					2			
			NY860	M14_nonT_NY860___MC			330					
				M14_nonT_NY860___TC					5			
		M14_rTreg	NY580	M14_rTreg_NY580___MC			1					
		M14_Tncl	BT580	M14_Tncl_BT580___MC			1					
	M16	M16_aTreg	BT580	M16_aTreg_BT580___MC			4			398	0.0276	
			BT860	M16_aTreg_BT860___MC			5					
			NY580	M16_aTreg_NY580___MC			7					
			NY860	M16_aTreg_NY860___MC			7					
		M16_nonT	BT580	M16_nonT_BT580___DC		3						
				M16_nonT_BT580___MC			56					
			BT860	M16_nonT_BT860___DC		2						
				M16_nonT_BT860___MC			100					
				M16_nonT_BT860___TC					5			
			NY580	M16_nonT_NY580___MC			80					
				M16_nonT_NY580___TC					1			
			NY860	M16_nonT_NY860___MC			126					
		M16_T8em	BT580	M16_T8em_BT580___MC			1					
		M16_T8em	NY860	M16_T8em_NY860___MC			1					

NK_cells	NK	NK_aTreg	BT580	NK_aTreg_BT580__NK				1		1394	0.1901	1394
				NK_aTreg_BT580__TC					1			
			NY580	NK_aTreg_NY580__TC					3			
			NY860	NK_aTreg_NY860__TC					1			
		NK_nonT	BT580	NK_nonT_BT580__MC			1					
				NK_nonT_BT580__NK				236				
				NK_nonT_BT580__TC					17			
			BT860	NK_nonT_BT860__BC	2							
				NK_nonT_BT860__NK				358				
				NK_nonT_BT860__TC					65			
			NY580	NK_nonT_NY580__MC			1					
				NK_nonT_NY580__NK				177				
				NK_nonT_NY580__TC					10			
			NY860	NK_nonT_NY860__NK				238				
				NK_nonT_NY860__TC					26			
		NK_T4em	NY860	NK_T4em_NY860__NK				1				
		NK_T4naive	NY860	NK_T4naive_NY860__TC					1			
		NK_T8em	BT580	NK_T8em_BT580__NK				15				
				NK_T8em_BT580__TC					25			
			BT860	NK_T8em_BT860__NK				39				
				NK_T8em_BT860__TC					47			
			NY580	NK_T8em_NY580__NK				13				
				NK_T8em_NY580__TC					5			
			NY860	NK_T8em_NY860__NK				41				
				NK_T8em_NY860__TC					27			
		NK_Tncl	BT580	NK_Tncl_BT580__NK				2				
				NK_Tncl_BT580__TC					8			
			BT860	NK_Tncl_BT860__NK				3				
				NK_Tncl_BT860__TC					7			
			NY580	NK_Tncl_NY580__NK				1				
				NK_Tncl_NY580__TC					10			
			NY860	NK_Tncl_NY860__NK				4				
				NK_Tncl_NY860__TC					8			
T_cells	aTreg	T_aTreg	BT580	T_aTreg_BT580__DC		1				921	0.0011	8326
				T_aTreg_BT580__TC					240			
			BT860	T_aTreg_BT860__TC					243			
			NY580	T_aTreg_NY580__TC					222			
			NY860	T_aTreg_NY860__TC					215			
	nonT	T_nonT	BT580	T_nonT_BT580__NK				50		426	0.4859	
				T_nonT_BT580__TC					46			
			BT860	T_nonT_BT860__NK				55				
				T_nonT_BT860__TC					79			
			NY580	T_nonT_NY580__NK				52				
				T_nonT_NY580__TC					33			
			NY860	T_nonT_NY860__NK				50				
				T_nonT_NY860__TC					61			
	rTreg	T_rTreg	BT580	T_rTreg_BT580__TC					313	1072	0.0028	
			BT860	T_rTreg_BT860__BC	1							
				T_rTreg_BT860__NK				1				
				T_rTreg_BT860__TC					232			
			NY580	T_rTreg_NY580__TC					337			
			NY860	T_rTreg_NY860__NK				1				
				T_rTreg_NY860__TC					187			
	T4em	T_T4em	BT580	T_T4em_BT580__MC			1			975	0.0092	
				T_T4em_BT580__NK				2				
				T_T4em_BT580__TC					327			
			BT860	T_T4em_BT860__NK				1				
				T_T4em_BT860__TC					258			
			NY580	T_T4em_NY580__NK				4				
				T_T4em_NY580__TC					250			
			NY860	T_T4em_NY860__NK				1				
				T_T4em_NY860__TC					131			
	T4naive	T_T4naive	BT580	T_T4naive_BT580__TC					482	1134	0.0009	
			BT860	T_T4naive_BT860__NK				1				
				T_T4naive_BT860__TC					264			
			NY580	T_T4naive_NY580__TC					291			
			NY860	T_T4naive_NY860__TC					96			
	T8em	T_T8em	BT580	T_T8em_BT580__NK				13		1031	0.0689	
				T_T8em_BT580__TC					253			
			BT860	T_T8em_BT860__NK				21				
				T_T8em_BT860__TC					283			
			NY580	T_T8em_NY580__NK				18				
				T_T8em_NY580__TC					248			
			NY860	T_T8em_NY860__NK				19				
				T_T8em_NY860__TC					176			
	T8naive	T_T8naive	BT580	T_T8naive_BT580__TC					318	1336	0.0000	
			BT860	T_T8naive_BT860__TC					486			
			NY580	T_T8naive_NY580__TC					256			
			NY860	T_T8naive_NY860__TC					276			
	Tncl	T_Tncl	BT580	T_Tncl_BT580__NK				9		1431	0.0231	
				T_Tncl_BT580__TC					192			
			BT860	T_Tncl_BT860__NK				6				
				T_Tncl_BT860__TC					360			
			NY580	T_Tncl_NY580__NK				7				
				T_Tncl_NY580__TC					372			
			NY860	T_Tncl_NY860__NK				11				
				T_Tncl_NY860__TC					474			
All (predicted)					**1521**	**155**	**1680**	**1518**	**8309**	**13183**		**13183**

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
2	10x (Clean)	BC	10085	85423	√	
		M14	2612		√	
		NK	8385		√	
		CD45RA+CD2	10479		√	
		T4	11213		√	
		CD45RA+T8n	11953		√	
		T8	10209		√	
		CD45RO+T4m	10224		√	
		CD4+CD25+T	10263		√	
	GEO (of R7)	M14_d1	425	30298	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T	435		√	
					√	
					√	
		T8_methano	4753		√	
		donor1_ IL-1	1247		√	
		donor2_ IL-1	1902		√	
		nonmalignan	4486		√	
		nonmalignan	3725		√	
		HLA-DR	48		√	
		HLA-DR_cont	2397		√	
		CD19	26		√	
		CD19_contro	1760		√	
		CD8	5662		√	
	BroadS1	Bn	1169	13183	√	
		Bm	491		√	
		DC	142		√	
		M14	1263		√	
		M16	398		√	
		NK	1394		√	
		aTreg	921		√	
		nonT	426		√	
		rTreg	1072		√	
		T4em	975		√	
		T4naive	1134		√	
		T8em	1031		√	
		T8naive	1336		√	
		Tncl	1431		√	
	BroadS2 (Clean)	BC	1884	12292		√
		DC	202			√
		pDC	68			√
		M14	1809			√
		M16	323			√
		NK	842			√
		T4	3380			√
		T8	3784			√

Accuracy:	0.929873				
Precision:	0.983778	0.96	0.94762757	0.69926393	0.941354
Recall/Ser	0.901274	0.355556	0.99296435	0.78978622	0.956728
Specificity	0.99731	0.999667	0.98848425	0.97502183	0.916732
F1_Score:	0.94072	0.518919	0.96976638	0.74177356	0.948979

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1698	0	10	2	174	1884
Dendritic_	19	96	89	1	65	270
Monocyte	0	2	2117	0	13	2132
NK_cells	2	0	0	665	175	842
T_cells	7	2	18	283	6854	7164
All	1726	100	2234	951	7281	12292

True/ Predicted						BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	BC	pbmc1	v2	A	pbmc1_v2_A_BC___BC	233					1884	0.0987	1884
					pbmc1_v2_A_BC___MC			3					
					pbmc1_v2_A_BC___NK				1				
					pbmc1_v2_A_BC___TC					51			
				B	pbmc1_v2_B_BC___BC	305							
					pbmc1_v2_B_BC___MC			2					
					pbmc1_v2_B_BC___NK				1				
					pbmc1_v2_B_BC___TC					80			
			v3		pbmc1_v3_BC___BC	316							
					pbmc1_v3_BC___TC					30			
		pbmc2	v2		pbmc2_V2_BC___BC	844							
					pbmc2_V2_BC___MC			5					
					pbmc2_V2_BC___TC					13			
Dendritic_cells	DC	pbmc1	v2	A	pbmc1_v2_A_DC___BC	1					202	0.5594	270
					pbmc1_v2_A_DC___DC		10						
					pbmc1_v2_A_DC___MC			34					
					pbmc1_v2_A_DC___TC					10			
				B	pbmc1_v2_B_DC___DC		10						
					pbmc1_v2_B_DC___MC			11					
					pbmc1_v2_B_DC___TC					12			
			v3		pbmc1_v3_DC___DC		13						
					pbmc1_v3_DC___MC			9					
					pbmc1_v3_DC___TC					16			
		pbmc2	v2		pbmc2_V2_DC___BC	1							
					pbmc2_V2_DC___DC		56						
					pbmc2_V2_DC___MC			10					
					pbmc2_V2_DC___TC					9			
	pDC	pbmc1	v2	A	pbmc1_v2_A_pDC___BC	8					68	0.8971	
					pbmc1_v2_A_pDC___DC		3						
					pbmc1_v2_A_pDC___MC			10					
					pbmc1_v2_A_pDC___TC					5			
				B	pbmc1_v2_B_pDC___MC			7					
					pbmc1_v2_B_pDC___TC					5			
		pbmc2	V2		pbmc2_V2_pDC___BC	9							
					pbmc2_V2_pDC___DC		4						
					pbmc2_V2_pDC___MC			8					
					pbmc2_V2_pDC___NK				1				
					pbmc2_V2_pDC___TC					8			
Monocytes	M14	pbmc1	v2	A	pbmc1_v2_A_M14___MC			637			1809	0.0044	2132
					pbmc1_v2_A_M14___TC					3			
				B	pbmc1_v2_B_M14___MC			378					
					pbmc1_v2_B_M14___TC					1			
			v3		pbmc1_v3_M14___MC			354					
		pbmc2	V2		pbmc2_V2_M14___DC		2						
					pbmc2_V2_M14___MC			432					
					pbmc2_V2_M14___TC					2			
	M16	pbmc1	v2	A	pbmc1_v2_A_M16___MC			95			323	0.0217	
					pbmc1_v2_A_M16___TC					7			
				B	pbmc1_v2_B_M16___MC			73					
			v3		pbmc1_v3_M16___MC			98					
		pbmc2	V2		pbmc2_V2_M16___MC			50					
NK_cells	NK	pbmc1	v2	A	pbmc1_v2_A_NK___BC	1					842	0.2102	842
					pbmc1_v2_A_NK___NK				123				
					pbmc1_v2_A_NK___TC					42			
				B	pbmc1_v2_B_NK___NK				175				
					pbmc1_v2_B_NK___TC					88			
			v3		pbmc1_v3_NK___NK				157				
					pbmc1_v3_NK___TC					37			
		pbmc2	V2		pbmc2_V2_NK___BC	1							
					pbmc2_V2_NK___NK				210				
					pbmc2_V2_NK___TC					8			
T_cells	T4	pbmc1	v2	A	pbmc1_v2_A_T4___BC	1					3380	0.0098	7164
					pbmc1_v2_A_T4___NK				4				
					pbmc1_v2_A_T4___TC					545			
				B	pbmc1_v2_B_T4___DC		1						
					pbmc1_v2_B_T4___MC			2					
					pbmc1_v2_B_T4___NK				4				
					pbmc1_v2_B_T4___TC					901			
			v3		pbmc1_v3_T4___NK				10				
					pbmc1_v3_T4___TC					950			
		pbmc2	V2		pbmc2_V2_T4___BC	3							
					pbmc2_V2_T4___DC		1						
					pbmc2_V2_T4___MC			5					
					pbmc2_V2_T4___NK				2				
					pbmc2_V2_T4___TC					951			
	T8	pbmc1	v2	A	pbmc1_v2_A_T8___MC			7			3784	0.0732	
					pbmc1_v2_A_T8___NK				84				
					pbmc1_v2_A_T8___TC					1083			
				B	pbmc1_v2_B_T8___MC			2					
					pbmc1_v2_B_T8___NK				51				
					pbmc1_v2_B_T8___TC					901			
			v3		pbmc1_v3_T8___NK				51				
					pbmc1_v3_T8___TC					911			
		pbmc2	V2		pbmc2_V2_T8___BC	3							
					pbmc2_V2_T8___MC			2					
					pbmc2_V2_T8___NK				77				
					pbmc2_V2_T8___TC					612			
All (predicted)						1726	100	2234	951	7281	12292		12292

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
3	10x (Clean)	BC	10085	85423		√
		M14	2612			√
		NK	8385			√
		CD45RA+CD25-T4naive	10479			√
		T4	11213			√
		CD45RA+T8naive	11953			√
		T8	10209			√
		CD45RO+T4mem	10224			√
		CD4+CD25+Treg	10263			√
	GEO (of R7)	M14_d1	425	30298	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T4	435		√	
					√	
					√	
		T8_methanol_SSC	4753		√	
		donor1_ IL-10-producing_Foxp3-_T4	1247		√	
		donor2_ IL-10-producing_Foxp3-_T4	1902		√	
		nonmalignant_P5_CD3+CD5intSSCint_T4	4486		√	
		nonmalignant_P5_CD3+CD5intSSCint_T4_afterther	3725		√	
		HLA-DR	48		√	
		HLA-DR_control	2397		√	
		CD19	26		√	
		CD19_control	1760		√	
		CD8	5662		√	
	BroadS1	Bn	1169	13183	√	
		Bm	491		√	
		DC	142		√	
		M14	1263		√	
		M16	398		√	
		NK	1394		√	
		aTreg	921		√	
		nonT	426		√	
		rTreg	1072		√	
		T4em	975		√	
		T4naive	1134		√	
		T8em	1031		√	
		T8naive	1336		√	
		Tncl	1431		√	
	BroadS2 (Clean)	BC	1884	12292	√	
		DC	202		√	
		pDC	68		√	
		M14	1809		√	
		M16	323		√	
		NK	842		√	
		T4	3380		√	
		T8	3784		√	

Accuracy:	0.198400899			
Precision:	0.46587537	0.02873688	1	0.914184
Recall/Sens	1.56E-02	0.75803982	0.020751	0.227491
Specificity:	0.99761077	0.19188272	1	0.934826
F1_Score:	0.03012857	0.05537455	0.040659	0.364322

Predicted	B_cells	Monocytes	NK_cells	T_cells	All
B_cells	157	9906	0	22	10085
Monocytes	6	1980	0	626	2612
NK_cells	1	7484	174	726	8385
T_cells	173	49531	0	14637	64341
All	337	68901	174	16011	85423

True/ Predicted			BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	BC	021-CD19+B___BC	157					10085	0.9844	10085
		021-CD19+B___MC			9906					
		021-CD19+B___TC					22			
Monocytes	M14	003-M14___BC	6					2612	0.2420	2612
		003-M14___MC			1980					
		003-M14___TC					626			
NK_cells	NK	018-CD56+NK___BC	1					8385	0.9792	8385
		018-CD56+NK___MC			7484					
		018-CD56+NK___NK				174				
		018-CD56+NK___TC					726			
T_cells	CD45RA+CD25-T4naive	025-CD4+CD45RA+CD25-NaiveT___BC	25					10479	0.8854	64341
		025-CD4+CD45RA+CD25-NaiveT___MC			9253					
		025-CD4+CD45RA+CD25-NaiveT___TC					1201			
	T4	026-T4___BC	59					11213	0.8352	
		026-T4___MC			9306					
		026-T4___TC					1848			
	CD45RA+T8naive	027-CD8+CD45RA+NaiveCytotoxicT___MC			11073			11953	0.9264	
		027-CD8+CD45RA+NaiveCytotoxicT___TC					880			
	T8	022-T8___BC	1					10209	0.6873	
		022-T8___MC			7016					
		022-T8___TC					3192			
	CD45RO+T4mem	023-CD4+CD45RO+MemoryT___MC			6889			10224	0.6738	
		023-CD4+CD45RO+MemoryT___TC					3335			
	CD4+CD25+Treg	024-CD4+CD25+RegulatoryT___BC	88					10263	0.5926	
		024-CD4+CD25+RegulatoryT___MC			5994					
		024-CD4+CD25+RegulatoryT___TC					4181			
All (predicted)			**337**	**0**	**68901**	**174**	**16011**	**85423**		**85423**

EXP	DataSets	Subtype	SubtypeN	TotalCell	Training	Testing
4	10x (Clean)	BC	10085	85423	√	
		M14	2612		√	
		NK	8385		√	
		CD45RA+CD25-T4naive	10479		√	
		T4	11213		√	
		CD45RA+T8naive	11953		√	
		T8	10209		√	
		CD45RO+T4mem	10224		√	
		CD4+CD25+Treg	10263		√	
	GEO (of R7)	M14_d1	425	30298		√
		M14_d2	431			√
		NK	309			√
		T4	222			√
		T8	310			√
		iNKT	325			√
		MAIT	382			√
		Vd1	284			√
		Vd2	204			√
		T4	965			√
		CCR5+CD69-T4	435			√
						√
						√
		T8_methanol_SSC	4753			√
		donor1_ IL-10-producing_Foxp3-_T4	1247			√
		donor2_ IL-10-producing_Foxp3-_T4	1902			√
		nonmalignant_P5_CD3+CD5intSSCint_T4	4486			√
		nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy	3725			√
		HLA-DR	48			√
		HLA-DR_control	2397			√
		CD19	26			√
		CD19_control	1760			√
		CD8	5662			√
	BroadS1	Bn	1169	13183	√	
		Bm	491		√	
		DC	142		√	
		M14	1263		√	
		M16	398		√	
		NK	1394		√	
		aTreg	921		√	
		nonT	426		√	
		rTreg	1072		√	
		T4em	975		√	
		T4naive	1134		√	
		T8em	1031		√	
		T8naive	1336		√	
		Tncl	1431		√	
	BroadS2 (Clean)	BC	1884	12292	√	
		DC	202		√	
		pDC	68		√	
		M14	1809		√	
		M16	323		√	
		NK	842		√	
		T4	3380		√	
		T8	3784		√	

Accuracy:	0.86065087				
Precision:	0.72616984	0	0.818458	0.105443	0.973685
Recall/Sensi	0.70380739	0	0.733414	0.996764	0.887077
Specificity:	0.98337542	0.999967	0.980109	0.912868	0.889362
F1_Score:	0.71481376	0	0.773606	0.190712	0.928366

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	1	204	79	245	1786
Monocytes	64	0	2421	465	351	3301
NK_cells	0	0	0	308	1	309
T_cells	410	0	333	2069	22090	24902
All	1731	1	2958	2921	22687	30298

True/ Predicted			BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	CD19_control	GEO_GSM3258348_CD19_control___BC	1249					1760	0.2903	1786
		GEO_GSM3258348_CD19_control___MC			197					
		GEO_GSM3258348_CD19_control___NK				79				
		GEO_GSM3258348_CD19_control___TC					235			
	CD19	GEO_GSM3258346_CD19___BC	8					26	0.6923	
		GEO_GSM3258346_CD19___DC		1						
		GEO_GSM3258346_CD19___MC			7					
		GEO_GSM3258346_CD19___TC					10			
Monocytes	M14_d1	GEO_GSM2773408_M14_d1___MC			420			425	0.0118	3301
		GEO_GSM2773408_M14_d1___NK				1				
		GEO_GSM2773408_M14_d1___TC					4			
	M14_d2	GEO_GSM2773409_M14_d2___BC	3					431	0.0278	
		GEO_GSM2773409_M14_d2___MC			419					
		GEO_GSM2773409_M14_d2___NK				4				
		GEO_GSM2773409_M14_d2___TC					5			
	HLA-DR	GEO_GSM3258345_HLA-DR_BC	5					48	0.3125	
		GEO_GSM3258345_HLA-DR_MC			33					
		GEO_GSM3258345_HLA-DR_NK				3				
		GEO_GSM3258345_HLA-DR_TC					7			
	HLA-DR_control	GEO_GSM3258347_HLA-DR_control_BC	56					2397	0.3538	
		GEO_GSM3258347_HLA-DR_control_MC			1549					
		GEO_GSM3258347_HLA-DR_control_NK				457				
		GEO_GSM3258347_HLA-DR_control_TC					335			
NK_cells	NK	GEO_GSM3544603_NK___NK				308		309	0.0032	309
		GEO_GSM3544603_NK___TC					1			
T_cells	T4	GEO_20190108_GSM3544603_T4___TC					222	222	0.0000	24902
	T8	GEO_20190108_GSM3544603_T8___MC			1			310	0.0161	
		GEO_20190108_GSM3544603_T8___NK				4				
		GEO_20190108_GSM3544603_T8___TC					305			
	iNKT	GEO_20190108_GSM3544603_iNKT___NK				37		325	0.1138	
		GEO_20190108_GSM3544603_iNKT___TC					288			
	MAIT	GEO_20190108_GSM3544603_MAIT___NK				20		382	0.0524	
		GEO_20190108_GSM3544603_MAIT___TC					362			
	Vd1	GEO_20190108_GSM3544603_Vd1___MC			1			284	0.4542	
		GEO_20190108_GSM3544603_Vd1___NK				128				
		GEO_20190108_GSM3544603_Vd1___TC					155			
	Vd2	GEO_20190108_GSM3544603_Vd2___NK				44		204	0.2157	
		GEO_20190108_GSM3544603_Vd2___TC					160			
	T4	GEO_20190620_GSM3209407_T4___NK				16		965	0.0166	
		GEO_20190620_GSM3209407_T4___TC					949			
	CCR5+CD69-T4	GEO_20190620_GSM3209408_CCR5+CD69-T4___NK				9		435	0.0207	
		GEO_20190620_GSM3209408_CCR5+CD69-T4___TC					426			
	T8_methanol_SSC	GEO_GSM3087629_T8_methanol_SSC_BC	183					4753	0.2981	
		GEO_GSM3087629_T8_methanol_SSC_MC			98					
		GEO_GSM3087629_T8_methanol_SSC_NK				1136				
		GEO_GSM3087629_T8_methanol_SSC_TC					3336			
	r1_ IL-10-producing_Foxp	GEO_GSM3430548_donor1_ IL-10-producing_Foxp3-_T4_NK				6		1247	0.0048	
		GEO_GSM3430548_donor1_ IL-10-producing_Foxp3-_T4_TC					1241			
	r2_ IL-10-producing_Foxp	GEO_GSM3430549_donor2_ IL-10-producing_Foxp3-_T4_BC	1					1902	0.0068	
		GEO_GSM3430549_donor2_ IL-10-producing_Foxp3-_T4_NK				12				
		GEO_GSM3430549_donor2_ IL-10-producing_Foxp3-_T4_TC					1889			
	lignant_P5_CD3+CD5intSS	GEO_GSM3478792_nonmalignant_P5_CD3+CD5intSSCint_T4_BC	1					4486	0.0069	
		GEO_GSM3478792_nonmalignant_P5_CD3+CD5intSSCint_T4_MC			22					
		GEO_GSM3478792_nonmalignant_P5_CD3+CD5intSSCint_T4_NK				8				
		GEO_GSM3478792_nonmalignant_P5_CD3+CD5intSSCint_T4_TC					4455			
	_P5_CD3+CD5intSSCint_T4	GEO_GSM3558027_nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy_BC	5					3725	0.0030	
		GEO_GSM3558027_nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy_NK				6				
		GEO_GSM3558027_nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy_TC					3714			
	CD8	GEO_GSM3087628_T8___BC	220					5662	0.1897	
		GEO_GSM3087628_T8___MC			211					
		GEO_GSM3087628_T8___NK				643				
		GEO_GSM3087628_T8___TC					4588			
All (predicted)			**1731**	**1**	**2958**	**2921**	**22687**	**30298**	**3.0906**	**30298**

SplitConfusionMatrix-R8

(Compared to R1 (R1 included ALL groups), R8 removed the 'EC', 'Other Tissue', and 'Dead Cells' groups.)

Train: 10x(Clean)+GEO(of R8)+BroadS2(Clean)

Test: BroadS1

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
1	10x (Clean)	BC	10085	85423	√	
		M14	2612		√	
		NK	8385		√	
		CD45RA+CD25-T4naive	10479		√	
		T4	11213		√	
		CD45RA+T8naive	11953		√	
		T8	10209		√	
		CD45RO+T4mem	10224		√	
		CD4+CD25+Treg	10263		√	
	GEO (of R8)	M14_d1	425	25545	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T4	435		√	
					√	
					√	
					√	
		donor1_ IL-10-producing_Foxp3-_T4	1247		√	
		donor2_ IL-10-producing_Foxp3-_T4	1902		√	
		nonmalignant_P5_CD3+CD5intSSCir	4486		√	
		nonmalignant_P5_CD3+CD5intSSCir	3725		√	
		HLA-DR	48		√	
		HLA-DR_control	2397		√	
		CD19	26		√	
		CD19_control	1760		√	
		CD8	5662		√	
	BroadS1	Bn	1169	13183		√
		Bm	491			√
		DC	142			√
		M14	1263			√
		M16	398			√
		NK	1394			√
		aTreg	921			√
		nonT	426			√
		rTreg	1072			√
		T4em	975			√
		T4naive	1134			√
		T8em	1031			√
		T8naive	1336			√
		Tncl	1431			√
	BroadS2 (Clean)	BC	1884	12292	√	
		DC	202		√	
		pDC	68		√	
		M14	1809		√	
		M16	323		√	
		NK	842		√	
		T4	3380		√	
		T8	3784		√	

Accuracy:	0.936509141					
Precision:	0.99605263	0.61111111	0.97329193	0.77738516	0.9538115	
Recall/Sensitivi	0.91204819	0.92957746	0.94340759	0.78909613	0.964809	
Specificity:	0.9994793	0.99355878	0.99626801	0.97328018	0.9199094	
F1_Score:	0.95220126	0.73743017	0.95811678	0.78319687	0.9592787	
Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1514	16	30	30	70	1660
Dendritic_cells	0	132	7	0	3	142
Monocytes	1	67	1567	0	26	1661
NK_cells	2	0	2	1100	290	1394
T_cells	3	1	4	285	8033	8326
All	1520	216	1610	1415	8422	13183

True/ Predicted					BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	Bn	Bn_aTreg	BT580	Bn_aTreg_BT580__BC	4					1169	0.0847	1660
			BT860	Bn_aTreg_BT860__BC	6							
			NY860	Bn_aTreg_NY860__BC	3							
		Bn_nonT	BT580	Bn_nonT_BT580__BC	235							
				Bn_nonT_BT580__DC		1						
				Bn_nonT_BT580__MC			2					
				Bn_nonT_BT580__NK				3				
				Bn_nonT_BT580__TC					6			
			BT860	Bn_nonT_BT860__BC	514							
				Bn_nonT_BT860__DC		4						
				Bn_nonT_BT860__MC			6					
				Bn_nonT_BT860__NK				12				
				Bn_nonT_BT860__TC					17			
			NY580	Bn_nonT_NY580__BC	143							
				Bn_nonT_NY580__DC		3						
				Bn_nonT_NY580__MC			7					
				Bn_nonT_NY580__NK				1				
				Bn_nonT_NY580__TC					10			
			NY860	Bn_nonT_NY860__BC	163							
				Bn_nonT_NY860__DC		3						
				Bn_nonT_NY860__MC			2					
				Bn_nonT_NY860__NK				6				
				Bn_nonT_NY860__TC					16			
		Bn_T4em	BT860	Bn_T4em_BT860__BC	1							
		Bn_Tncl	BT860	Bn_Tncl_BT860__BC	1							
	Bm	Bm_aTreg	BT860	Bm_aTreg_BT860__BC	6					491	0.0957	
			NY580	Bm_aTreg_NY580__BC	1							
			NY860	Bm_aTreg_NY860__BC	2							
		Bm_nonT	BT580	Bm_nonT_BT580__BC	85							
				Bm_nonT_BT580__MC			3					
			BT860	Bm_nonT_BT860__BC	206							
				Bm_nonT_BT860__DC		2						
				Bm_nonT_BT860__MC			7					
				Bm_nonT_BT860__TC					8			
			NY580	Bm_nonT_NY580__BC	59							
				Bm_nonT_NY580__DC		1						
				Bm_nonT_NY580__TC					3			
			NY860	Bm_nonT_NY860__BC	85							
				Bm_nonT_NY860__DC		2						
				Bm_nonT_NY860__MC			3					
				Bm_nonT_NY860__NK				8				
				Bm_nonT_NY860__TC					10			
Dendritic_cells	DC	DC_aTreg	BT860	DC_aTreg_BT860__DC		1				142	0.0704	142
			NY580	DC_aTreg_NY580__DC		1						
		DC_nonT	BT580	DC_nonT_BT580__DC		50						
				DC_nonT_BT580__MC			3					
				DC_nonT_BT580__TC					1			
			BT860	DC_nonT_BT860__DC		18						
				DC_nonT_BT860__TC					1			
			NY580	DC_nonT_NY580__DC		45						
				DC_nonT_NY580__MC			1					
			NY860	DC_nonT_NY860__DC		17						
				DC_nonT_NY860__MC			3					
				DC_nonT_NY860__TC					1			
Monocytes	M14	M14_aTreg	BT580	M14_aTreg_BT580__MC			1			1263	0.0530	1661
			BT860	M14_aTreg_BT860__DC		1						
				M14_aTreg_BT860__MC			3					
			NY580	M14_aTreg_NY580__MC			2					
			NY860	M14_aTreg_NY860__MC			2					
		M14_nonT	BT580	M14_nonT_BT580__DC		10						
				M14_nonT_BT580__MC			226					
				M14_nonT_BT580__TC					2			
			BT860	M14_nonT_BT860__DC		13						
				M14_nonT_BT860__MC			319					
				M14_nonT_BT860__TC					6			
			NY580	M14_nonT_NY580__DC		11						
				M14_nonT_NY580__MC			327					
				M14_nonT_NY580__TC					3			
			NY860	M14_nonT_NY860__DC		13						
				M14_nonT_NY860__MC			314					
				M14_nonT_NY860__TC					8			
		M14_rTreg	NY580	M14_rTreg_NY580__MC			1					
		M14_Tncl	BT580	M14_Tncl_BT580__MC			1					
	M16	M16_aTreg	BT580	M16_aTreg_BT580__MC			4			398	0.0678	
			BT860	M16_aTreg_BT860__DC		1						
				M16_aTreg_BT860__MC			4					
			NY580	M16_aTreg_NY580__DC		1						
				M16_aTreg_NY580__MC			6					
			NY860	M16_aTreg_NY860__MC			7					
		M16_nonT	BT580	M16_nonT_BT580__DC		7						
				M16_nonT_BT580__MC			52					
			BT860	M16_nonT_BT860__BC	1							
				M16_nonT_BT860__DC		4						
				M16_nonT_BT860__MC			97					
				M16_nonT_BT860__TC					5			
			NY580	M16_nonT_NY580__DC		3						
				M16_nonT_NY580__MC			77					
				M16_nonT_NY580__TC					1			
			NY860	M16_nonT_NY860__DC		3						
				M16_nonT_NY860__MC			122					
				M16_nonT_NY860__TC					1			
		M16_T8em	BT580	M16_T8em_BT580__MC			1					
		M16_T8em	NY860	M16_T8em_NY860__MC			1					

NK_cells	NK	NK_aTreg	BT580	NK_aTreg_BT580___TC					2	1394	0.2109	1394
			NY580	NK_aTreg_NY580___TC					3			
			NY860	NK_aTreg_NY860___TC					1			
		NK_nonT	BT580	NK_nonT_BT580___MC			1					
				NK_nonT_BT580___NK				232				
				NK_nonT_BT580___TC					21			
			BT860	NK_nonT_BT860___BC	2							
				NK_nonT_BT860___NK				359				
				NK_nonT_BT860___TC					64			
			NY580	NK_nonT_NY580___MC			1					
				NK_nonT_NY580___NK				162				
				NK_nonT_NY580___TC					25			
			NY860	NK_nonT_NY860___NK				231				
				NK_nonT_NY860___TC					33			
		NK_T4em	NY860	NK_T4em_NY860___NK				1				
		NK_T4naive	NY860	NK_T4naive_NY860___TC					1			
		NK_T8em	BT580	NK_T8em_BT580___NK				18				
				NK_T8em_BT580___TC					22			
			BT860	NK_T8em_BT860___NK				37				
				NK_T8em_BT860___TC					49			
			NY580	NK_T8em_NY580___NK				13				
				NK_T8em_NY580___TC					5			
			NY860	NK_T8em_NY860___NK				36				
				NK_T8em_NY860___TC					32			
		NK_Tncl	BT580	NK_Tncl_BT580___NK				3				
				NK_Tncl_BT580___TC					7			
			BT860	NK_Tncl_BT860___NK				3				
				NK_Tncl_BT860___TC					7			
			NY580	NK_Tncl_NY580___NK				1				
				NK_Tncl_NY580___TC					10			
			NY860	NK_Tncl_NY860___NK				4				
				NK_Tncl_NY860___TC					8			
T_cells	aTreg	T_aTreg	BT580	T_aTreg_BT580___DC		1				921	0.0022	8326
				T_aTreg_BT580___TC					240			
			BT860	T_aTreg_BT860___NK				1				
				T_aTreg_BT860___TC					242			
			NY580	T_aTreg_NY580___TC					222			
			NY860	T_aTreg_NY860___TC					215			
	nonT	T_nonT	BT580	T_nonT_BT580___NK				47		426	0.4484	
				T_nonT_BT580___TC					49			
			BT860	T_nonT_BT860___NK				56				
				T_nonT_BT860___TC					78			
			NY580	T_nonT_NY580___NK				45				
				T_nonT_NY580___TC					40			
			NY860	T_nonT_NY860___NK				43				
				T_nonT_NY860___TC					68			
	rTreg	T_rTreg	BT580	T_rTreg_BT580___MC			1			1072	0.0037	
				T_rTreg_BT580___NK				2				
				T_rTreg_BT580___TC					310			
			BT860	T_rTreg_BT860___BC	1							
				T_rTreg_BT860___TC					233			
			NY580	T_rTreg_NY580___TC					337			
			NY860	T_rTreg_NY860___TC					188			
	T4em	T_T4em	BT580	T_T4em_BT580___MC			2			975	0.0092	
				T_T4em_BT580___NK				5				
				T_T4em_BT580___TC					323			
			BT860	T_T4em_BT860___TC					259			
			NY580	T_T4em_NY580___NK				1				
				T_T4em_NY580___TC					253			
			NY860	T_T4em_NY860___NK				1				
				T_T4em_NY860___TC					131			
	T4naive	T_T4naive	BT580	T_T4naive_BT580___BC	2					1134	0.0026	
				T_T4naive_BT580___NK				1				
				T_T4naive_BT580___TC					479			
			BT860	T_T4naive_BT860___TC					265			
			NY580	T_T4naive_NY580___TC					291			
			NY860	T_T4naive_NY860___TC					96			
	T8em	T_T8em	BT580	T_T8em_BT580___NK				12		1031	0.0533	
				T_T8em_BT580___TC					254			
			BT860	T_T8em_BT860___NK				19				
				T_T8em_BT860___TC					285			
			NY580	T_T8em_NY580___NK				11				
				T_T8em_NY580___TC					255			
			NY860	T_T8em_NY860___NK				13				
				T_T8em_NY860___TC					182			
	T8naive	T_T8naive	BT580	T_T8naive_BT580___TC					318	1336	0.0000	
			BT860	T_T8naive_BT860___TC					486			
			NY580	T_T8naive_NY580___TC					256			
			NY860	T_T8naive_NY860___TC					276			
	Tncl	T_Tncl	BT580	T_Tncl_BT580___MC			1			1431	0.0203	
				T_Tncl_BT580___NK				5				
				T_Tncl_BT580___TC					195			
			BT860	T_Tncl_BT860___NK				6				
				T_Tncl_BT860___TC					360			
			NY580	T_Tncl_NY580___NK				8				
				T_Tncl_NY580___TC					371			
			NY860	T_Tncl_NY860___NK				9				
				T_Tncl_NY860___TC					476			
All (predicted)					**1520**	**216**	**1610**	**1415**	**8422**	**13183**		**13183**

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
2	10x (Clean)	BC	10085	85423	√	
		M14	2612		√	
		NK	8385		√	
		CD45RA+CD2	10479		√	
		T4	11213		√	
		CD45RA+T8n	11953		√	
		T8	10209		√	
		CD45RO+T4m	10224		√	
		CD4+CD25+T	10263		√	
	GEO (of R8)	M14_d1	425	25545	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T	435		√	
					√	
					√	
					√	
		donor1_ IL-1(	1247		√	
		donor2_ IL-1(	1902		√	
		nonmalignar	4486		√	
		nonmalignar	3725		√	
		HLA-DR	48		√	
		HLA-DR_cont	2397		√	
		CD19	26		√	
		CD19_contro	1760		√	
		CD8	5662		√	
	BroadS1	Bn	1169	13183	√	
		Bm	491		√	
		DC	142		√	
		M14	1263		√	
		M16	398		√	
		NK	1394		√	
		aTreg	921		√	
		nonT	426		√	
		rTreg	1072		√	
		T4em	975		√	
		T4naive	1134		√	
		T8em	1031		√	
		T8naive	1336		√	
		Tncl	1431		√	
	BroadS2 (Clean)	BC	1884	12292		√
		DC	202			√
		pDC	68			√
		M14	1809			√
		M16	323			√
		NK	842			√
		T4	3380			√
		T8	3784			√

Accuracy:	0.913358				
Precision:	0.964365	0.886792	0.87063228	0.61978221	0.961799
Recall/Ser	0.919321	0.174074	0.98170732	0.8111639	0.931323
Specificity	0.993851	0.999501	0.96938976	0.96340611	0.948323
F1_Score:	0.941304	0.291022	0.92283951	0.7026749	0.946316

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1732	0	51	0	101	1884
Dendritic_	8	47	204	0	11	270
Monocyte	26	3	2093	0	10	2132
NK_cells	3	0	13	683	143	842
T_cells	27	3	43	419	6672	7164
All	1796	53	2404	1102	6937	12292

True/ Predicted						BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	BC	pbmc1	v2	A	pbmc1_v2_A_BC___BC	239					1884	0.0807	1884
					pbmc1_v2_A_BC___MC			10					
					pbmc1_v2_A_BC___TC					39			
				B	pbmc1_v2_B_BC___BC	350							
					pbmc1_v2_B_BC___MC			13					
					pbmc1_v2_B_BC___TC					25			
			v3		pbmc1_v3_BC___BC	307							
					pbmc1_v3_BC___MC			15					
					pbmc1_v3_BC___TC					24			
		pbmc2	v2		pbmc2_V2_BC___BC	836							
					pbmc2_V2_BC___MC			13					
					pbmc2_V2_BC___TC					13			
Dendritic_cells	DC	pbmc1	v2	A	pbmc1_v2_A_DC___DC		6				202	0.7673	270
					pbmc1_v2_A_DC___MC			48					
					pbmc1_v2_A_DC___TC					1			
				B	pbmc1_v2_B_DC___MC			33					
			v3		pbmc1_v3_DC___DC		2						
					pbmc1_v3_DC___MC			32					
					pbmc1_v3_DC___TC					4			
		pbmc2	v2		pbmc2_V2_DC___DC		39						
					pbmc2_V2_DC___MC			34					
					pbmc2_V2_DC___TC					3			
	pDC	pbmc1	v2	A	pbmc1_v2_A_pDC___BC	3					68	1.0000	
					pbmc1_v2_A_pDC___MC			21					
					pbmc1_v2_A_pDC___TC					2			
				B	pbmc1_v2_B_pDC___MC			12					
		pbmc2	V2		pbmc2_V2_pDC___BC	5							
					pbmc2_V2_pDC___MC			24					
					pbmc2_V2_pDC___TC					1			
Monocytes	M14	pbmc1	v2	A	pbmc1_v2_A_M14___BC	16					1809	0.0182	2132
					pbmc1_v2_A_M14___DC		2						
					pbmc1_v2_A_M14___MC			619					
					pbmc1_v2_A_M14___TC					3			
				B	pbmc1_v2_B_M14___BC	1							
					pbmc1_v2_B_M14___MC			376					
					pbmc1_v2_B_M14___TC					2			
			v3		pbmc1_v3_M14___MC			354					
		pbmc2	V2		pbmc2_V2_M14___BC	6							
					pbmc2_V2_M14___DC		1						
					pbmc2_V2_M14___MC			427					
					pbmc2_V2_M14___TC					2			
	M16	pbmc1	v2	A	pbmc1_v2_A_M16___BC	3					323	0.0186	
					pbmc1_v2_A_M16___MC			96					
					pbmc1_v2_A_M16___TC					3			
				B	pbmc1_v2_B_M16___MC			73					
			v3		pbmc1_v3_M16___MC			98					
		pbmc2	V2		pbmc2_V2_M16___MC			50					
NK_cells	NK	pbmc1	v2	A	pbmc1_v2_A_NK___MC			5			842	0.1888	842
					pbmc1_v2_A_NK___NK				122				
					pbmc1_v2_A_NK___TC					39			
				B	pbmc1_v2_B_NK___BC	1							
					pbmc1_v2_B_NK___MC			5					
					pbmc1_v2_B_NK___NK				180				
					pbmc1_v2_B_NK___TC					77			
			v3		pbmc1_v3_NK___MC			3					
					pbmc1_v3_NK___NK				169				
					pbmc1_v3_NK___TC					22			
		pbmc2	V2		pbmc2_V2_NK___BC	2							
					pbmc2_V2_NK___NK				212				
					pbmc2_V2_NK___TC					5			
T_cells	T4	pbmc1	v2	A	pbmc1_v2_A_T4___BC	2					3380	0.0157	7164
					pbmc1_v2_A_T4___MC			3					
					pbmc1_v2_A_T4___NK				5				
					pbmc1_v2_A_T4___TC					540			
				B	pbmc1_v2_B_T4___BC	1							
					pbmc1_v2_B_T4___MC			4					
					pbmc1_v2_B_T4___NK				6				
					pbmc1_v2_B_T4___TC					897			
			v3		pbmc1_v3_T4___MC			3					
					pbmc1_v3_T4___NK				9				
					pbmc1_v3_T4___TC					948			
		pbmc2	V2		pbmc2_V2_T4___BC	8							
					pbmc2_V2_T4___DC		3						
					pbmc2_V2_T4___MC			5					
					pbmc2_V2_T4___NK				4				
					pbmc2_V2_T4___TC					942			
	T8	pbmc1	v2	A	pbmc1_v2_A_T8___BC	9					3784	0.1160	
					pbmc1_v2_A_T8___MC			18					
					pbmc1_v2_A_T8___NK				100				
					pbmc1_v2_A_T8___TC					1047			
				B	pbmc1_v2_B_T8___MC			5					
					pbmc1_v2_B_T8___NK				77				
					pbmc1_v2_B_T8___TC					872			
			v3		pbmc1_v3_T8___BC	1							
					pbmc1_v3_T8___MC			1					
					pbmc1_v3_T8___NK				96				
					pbmc1_v3_T8___TC					864			
		pbmc2	V2		pbmc2_V2_T8___BC	6							
					pbmc2_V2_T8___MC			4					
					pbmc2_V2_T8___NK				122				
					pbmc2_V2_T8___TC					562			
All (predicted)						1796	53	2404	1102	6937	12292		12292

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
3	10x (Clean)	BC	10085	85423		√
		M14	2612			√
		NK	8385			√
		CD45RA+CD25-T4naive	10479			√
		T4	11213			√
		CD45RA+T8naive	11953			√
		T8	10209			√
		CD45RO+T4mem	10224			√
		CD4+CD25+Treg	10263			√
	GEO (of R8)	M14_d1	425	25545	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T4	435		√	
					√	
					√	
					√	
		donor1_ IL-10-producing_Foxp3-_T4	1247		√	
		donor2_ IL-10-producing_Foxp3-_T4	1902		√	
		nonmalignant_P5_CD3+CD5intSSCint_T4	4486		√	
		nonmalignant_P5_CD3+CD5intSSCint_T4_afterthe	3725		√	
		HLA-DR	48		√	
		HLA-DR_control	2397		√	
		CD19	26		√	
		CD19_control	1760		√	
		CD8	5662		√	
	BroadS1	Bn	1169	13183	√	
		Bm	491		√	
		DC	142		√	
		M14	1263		√	
		M16	398		√	
		NK	1394		√	
		aTreg	921		√	
		nonT	426		√	
		rTreg	1072		√	
		T4em	975		√	
		T4naive	1134		√	
		T8em	1031		√	
		T8naive	1336		√	
		Tncl	1431		√	
	BroadS2 (Clean)	BC	1884	12292	√	
		DC	202		√	
		pDC	68		√	
		M14	1809		√	
		M16	323		√	
		NK	842		√	
		T4	3380		√	
		T8	3784		√	

Accuracy:	0.127588589				
Precision:	0.83802817	0	0.031333	0.833333	0.949289
Recall/Sens	3.54E-02	0	0.916539	0.000596	1.27E-01
Specificity:	0.99908413	0.99989464	0.106278	0.999987	0.979366
F1_Score:	0.06792884	0	0.060595	0.001192	0.223344

Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	357	1	9720	0	7	10085
Monocytes	25	8	2394	1	184	2612
NK_cells	24	0	8112	5	244	8385
T_cells	20	0	56178	0	8143	64341
All	426	9	76404	6	8578	85423

True/ Predicted			BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	BC	021-CD19+B___BC	357					10085	0.9646	10085
		021-CD19+B___DC		1						
		021-CD19+B___MC			9720					
		021-CD19+B___TC					7			
Monocytes	M14	003-M14___BC	25					2612	0.0835	2612
		003-M14___DC		8						
		003-M14___MC			2394					
		003-M14___NK				1				
		003-M14___TC					184			
NK_cells	NK	018-CD56+NK___BC	24					8385	0.9994	8385
		018-CD56+NK___MC			8112					
		018-CD56+NK___NK				5				
		018-CD56+NK___TC					244			
T_cells	CD45RA+CD25-T4naive	025-CD4+CD45RA+CD25-NaiveT___BC	10					10479	0.9295	64341
		025-CD4+CD45RA+CD25-NaiveT___MC			9730					
		025-CD4+CD45RA+CD25-NaiveT___TC					739			
	T4	026-T4___BC	3					11213	0.9108	
		026-T4___MC			10210					
		026-T4___TC					1000			
	CD45RA+T8naive	027-CD8+CD45RA+NaiveCytotoxicT___MC			11452			11953	0.9581	
		027-CD8+CD45RA+NaiveCytotoxicT___TC					501			
	T8	022-T8___BC	1					10209	0.8456	
		022-T8___MC			8632					
		022-T8___TC					1576			
	CD45RO+T4mem	023-CD4+CD45RO+MemoryT___MC			8309			10224	0.8127	
		023-CD4+CD45RO+MemoryT___TC					1915			
	CD4+CD25+Treg	024-CD4+CD25+RegulatoryT___BC	6					10263	0.7650	
		024-CD4+CD25+RegulatoryT___MC			7845					
		024-CD4+CD25+RegulatoryT___TC					2412			
All (predicted)			**426**	**9**	**76404**	**6**	**8578**	**85423**		**85423**

EXP	DataSets	Subtype	SubtypeN	TotalCell	Training	Testing
4	10x (Clean)	BC	10085	85423	√	
		M14	2612		√	
		NK	8385		√	
		CD45RA+CD25-T4naive	10479		√	
		T4	11213		√	
		CD45RA+T8naive	11953		√	
		T8	10209		√	
		CD45RO+T4mem	10224		√	
		CD4+CD25+Treg	10263		√	
	GEO (of R8)	M14_d1	425	25545		√
		M14_d2	431			√
		NK	309			√
		T4	222			√
		T8	310			√
		iNKT	325			√
		MAIT	382			√
		Vd1	284			√
		Vd2	204			√
		T4	965			√
		CCR5+CD69-T4	435			√
						√
						√
						√
		donor1_ IL-10-producing_Foxp3-_T4	1247			√
		donor2_ IL-10-producing_Foxp3-_T4	1902			√
		nonmalignant_P5_CD3+CD5intSSCint_T4	4486			√
		nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy	3725			√
		HLA-DR	48			√
		HLA-DR_control	2397			√
		CD19	26			√
		CD19_control	1760			√
		CD8	5662			√
	BroadS1	Bn	1169	13183	√	
		Bm	491		√	
		DC	142		√	
		M14	1263		√	
		M16	398		√	
		NK	1394		√	
		aTreg	921		√	
		nonT	426		√	
		rTreg	1072		√	
		T4em	975		√	
		T4naive	1134		√	
		T8em	1031		√	
		T8naive	1336		√	
		Tncl	1431		√	
	BroadS2 (Clean)	BC	1884	12292	√	
		DC	202		√	
		pDC	68		√	
		M14	1809		√	
		M16	323		√	
		NK	842		√	
		T4	3380		√	
		T8	3784		√	

Accuracy:	0.89019378				
Precision:	0.8120155	0	0.846504	0.172549	0.969149
Recall/Sensi	0.70380739	0	0.733414	0.996764	0.930766
Specificity:	0.98775201	0.999961	0.980264	0.941473	0.889362
F1_Score:	0.75404919	0	0.785911	0.294174	0.94957

Predicted	B_cells	ritic_cells	onocytes	NK_cells	T_cells	All
B_cells	1257	1	204	79	245	1786
Monocytes	64	0	2421	465	351	3301
NK_cells	0	0	0	308	1	309
T_cells	227	0	235	933	18754	20149
All	1548	1	2860	1785	19351	25545

True/ Predicted			BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	CD19_control	GEO_GSM3258348_CD19_control___BC	1249					1760	0.2903	1786
		GEO_GSM3258348_CD19_control___MC			197					
		GEO_GSM3258348_CD19_control___NK				79				
		GEO_GSM3258348_CD19_control___TC					235			
	CD19	GEO_GSM3258346_CD19___BC	8					26	0.6923	
		GEO_GSM3258346_CD19___DC		1						
		GEO_GSM3258346_CD19___MC			7					
		GEO_GSM3258346_CD19___TC					10			
Monocytes	M14_d1	GEO_GSM2773408_M14_d1___MC			420			425	0.0118	3301
		GEO_GSM2773408_M14_d1___NK				1				
		GEO_GSM2773408_M14_d1___TC					4			
	M14_d2	GEO_GSM2773409_M14_d2___BC	3					431	0.0278	
		GEO_GSM2773409_M14_d2___MC			419					
		GEO_GSM2773409_M14_d2___NK				4				
		GEO_GSM2773409_M14_d2___TC					5			
	HLA-DR	GEO_GSM3258345_HLA-DR_BC	5					48	0.3125	
		GEO_GSM3258345_HLA-DR_MC			33					
		GEO_GSM3258345_HLA-DR_NK				3				
		GEO_GSM3258345_HLA-DR_TC					7			
	HLA-DR_control	GEO_GSM3258347_HLA-DR_control_BC	56					2397	0.3538	
		GEO_GSM3258347_HLA-DR_control_MC			1549					
		GEO_GSM3258347_HLA-DR_control_NK				457				
		GEO_GSM3258347_HLA-DR_control_TC					335			
NK_cells	NK	GEO_GSM3544603_NK___NK				308		309	0.0032	309
		GEO_GSM3544603_NK___TC					1			
T_cells	T4	GEO_20190108_GSM3544603_T4___TC					222	222	0.0000	20149
	T8	GEO_20190108_GSM3544603_T8___MC			1			310	0.0161	
		GEO_20190108_GSM3544603_T8___NK				4				
		GEO_20190108_GSM3544603_T8___TC					305			
	iNKT	GEO_20190108_GSM3544603_iNKT___NK				37		325	0.1138	
		GEO_20190108_GSM3544603_iNKT___TC					288			
	MAIT	GEO_20190108_GSM3544603_MAIT___NK				20		382	0.0524	
		GEO_20190108_GSM3544603_MAIT___TC					362			
	Vd1	GEO_20190108_GSM3544603_Vd1___MC			1			284	0.4542	
		GEO_20190108_GSM3544603_Vd1___NK				128				
		GEO_20190108_GSM3544603_Vd1___TC					155			
	Vd2	GEO_20190108_GSM3544603_Vd2___NK				44		204	0.2157	
		GEO_20190108_GSM3544603_Vd2___TC					160			
	T4	GEO_20190620_GSM3209407_T4___NK				16		965	0.0166	
		GEO_20190620_GSM3209407_T4___TC					949			
	CCR5+CD69-T4	GEO_20190620_GSM3209408_CCR5+CD69-T4___NK				9		435	0.0207	
		GEO_20190620_GSM3209408_CCR5+CD69-T4___TC					426			
	r1_ IL-10-producing_Foxp3	GEO_GSM3430548_donor1_ IL-10-producing_Foxp3-_T4_NK				6		1247	0.0048	
		GEO_GSM3430548_donor1_ IL-10-producing_Foxp3-_T4_TC					1241			
	r2_ IL-10-producing_Foxp3	GEO_GSM3430549_donor2_ IL-10-producing_Foxp3-_T4_BC	1					1902	0.0068	
		GEO_GSM3430549_donor2_ IL-10-producing_Foxp3-_T4_NK				12				
		GEO_GSM3430549_donor2_ IL-10-producing_Foxp3-_T4_TC					1889			
	lignant_P5_CD3+CD5intSS	GEO_GSM3478792_nonmalignant_P5_CD3+CD5intSSCint_T4_BC	1					4486	0.0069	
		GEO_GSM3478792_nonmalignant_P5_CD3+CD5intSSCint_T4_MC			22					
		GEO_GSM3478792_nonmalignant_P5_CD3+CD5intSSCint_T4_NK				8				
		GEO_GSM3478792_nonmalignant_P5_CD3+CD5intSSCint_T4_TC					4455			
	_P5_CD3+CD5intSSCint_T4	GEO_GSM3558027_nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy_BC	5					3725	0.0030	
		GEO_GSM3558027_nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy_NK				6				
		GEO_GSM3558027_nonmalignant_P5_CD3+CD5intSSCint_T4_aftertherapy_TC					3714			
	CD8	GEO_GSM3087628_T8___BC	220					5662	0.1897	
		GEO_GSM3087628_T8___MC			211					
		GEO_GSM3087628_T8___NK				643				
		GEO_GSM3087628_T8___TC					4588			
All (predicted)			**1548**	**1**	**2860**	**1785**	**19351**	**25545**	**2.7925**	**25545**

SplitConfusionMatrix-R12

(Compared to R1 (R1 included ALL groups), R12 removed the 'EC', 'Other Tissue', 'Dead Cells', and 'Activated Cells' groups.)

Train: 10x(Clean)+GEO(of R12)+BroadS2(Clean)

Test: BroadS1

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
1	10x (Clean)	BC	10085	85423	√	
		M14	2612		√	
		NK	8385		√	
		CD45RA+CD25-T4naive	10479		√	
		T4	11213		√	
		CD45RA+T8naive	11953		√	
		T8	10209		√	
		CD45RO+T4mem	10224		√	
		CD4+CD25+Treg	10263		√	
	GEO (of R12)	M14_d1	425	14185	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T4	435		√	
					√	
					√	
					√	
					√	
					√	
					√	
					√	
		HLA-DR	48		√	
		HLA-DR_control	2397		√	
		CD19	26		√	
		CD19_control	1760		√	
		CD8	5662		√	
	BroadS1	Bn	1169	13183		√
		Bm	491			√
		DC	142			√
		M14	1263			√
		M16	398			√
		NK	1394			√
		aTreg	921			√
		nonT	426			√
		rTreg	1072			√
		T4em	975			√
		T4naive	1134			√
		T8em	1031			√
		T8naive	1336			√
		Tncl	1431			√
	BroadS2 (Clean)	BC	1884	12292	√	
		DC	202		√	
		pDC	68		√	
		M14	1809		√	
		M16	323		√	
		NK	842		√	
		T4	3380		√	
		T8	3784		√	

Accuracy:	0.938253812				
Precision:	0.99416721	0.79518072	0.95283019	0.78799392	0.9513118
Recall/Sensitivi	0.92409639	0.92957746	0.97290789	0.74390244	0.9668508
Specificity:	0.99921895	0.99739284	0.99305676	0.97633387	0.915174
F1_Score:	0.95785201	0.85714286	0.96276437	0.76531365	0.9590184

Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1534	24	53	16	33	1660
Dendritic_cells	0	132	9	0	1	142
Monocytes	7	9	1616	0	29	1661
NK_cells	2	0	6	1037	349	1394
T_cells	0	1	12	263	8050	8326
All	1543	166	1696	1316	8462	13183

True/ Predicted					BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	Bn	Bn_aTreg	BT580	Bn_aTreg_BT580__BC	4					1169	0.0719	1660
			BT860	Bn_aTreg_BT860__BC	6							
			NY860	Bn_aTreg_NY860__BC	2							
				Bn_aTreg_NY860__MC			1					
		Bn_nonT	BT580	Bn_nonT_BT580__BC	235							
				Bn_nonT_BT580__DC		3						
				Bn_nonT_BT580__MC			6					
				Bn_nonT_BT580__NK				1				
				Bn_nonT_BT580__TC					2			
			BT860	Bn_nonT_BT860__BC	518							
				Bn_nonT_BT860__DC		7						
				Bn_nonT_BT860__MC			21					
				Bn_nonT_BT860__NK				3				
				Bn_nonT_BT860__TC					4			
			NY580	Bn_nonT_NY580__BC	150							
				Bn_nonT_NY580__DC		3						
				Bn_nonT_NY580__MC			3					
				Bn_nonT_NY580__NK				1				
				Bn_nonT_NY580__TC					7			
			NY860	Bn_nonT_NY860__BC	168							
				Bn_nonT_NY860__DC		5						
				Bn_nonT_NY860__MC			4					
				Bn_nonT_NY860__NK				6				
				Bn_nonT_NY860__TC					7			
		Bn_T4em	BT860	Bn_T4em_BT860__BC	1							
		Bn_Tncl	BT860	Bn_Tncl_BT860__BC	1							
	Bm	Bm_aTreg	BT860	Bm_aTreg_BT860__BC	6					491	0.0855	
			NY580	Bm_aTreg_NY580__BC	1							
			NY860	Bm_aTreg_NY860__BC	2							
		Bm_nonT	BT580	Bm_nonT_BT580__BC	86							
				Bm_nonT_BT580__MC			2					
			BT860	Bm_nonT_BT860__BC	209							
				Bm_nonT_BT860__DC		2						
				Bm_nonT_BT860__MC			8					
				Bm_nonT_BT860__TC					4			
			NY580	Bm_nonT_NY580__BC	58							
				Bm_nonT_NY580__DC		1						
				Bm_nonT_NY580__MC			1					
				Bm_nonT_NY580__TC					3			
			NY860	Bm_nonT_NY860__BC	87							
				Bm_nonT_NY860__DC		3						
				Bm_nonT_NY860__MC			7					
				Bm_nonT_NY860__NK				5				
				Bm_nonT_NY860__TC					6			
Dendritic_cells	DC	DC_aTreg	BT860	DC_aTreg_BT860__DC		1				142	0.0704	142
			NY580	DC_aTreg_NY580__DC		1						
		DC_nonT	BT580	DC_nonT_BT580__DC		51						
				DC_nonT_BT580__MC			3					
			BT860	DC_nonT_BT860__DC		19						
			NY580	DC_nonT_NY580__DC		45						
				DC_nonT_NY580__MC			1					
			NY860	DC_nonT_NY860__DC		15						
				DC_nonT_NY860__MC			5					
				DC_nonT_NY860__TC					1			
Monocytes	M14	M14_aTreg	BT580	M14_aTreg_BT580__MC			1			1263	0.0245	1661
			BT860	M14_aTreg_BT860__MC			4					
			NY580	M14_aTreg_NY580__MC			2					
			NY860	M14_aTreg_NY860__MC			2					
		M14_nonT	BT580	M14_nonT_BT580__BC	3							
				M14_nonT_BT580__DC		1						
				M14_nonT_BT580__MC			231					
				M14_nonT_BT580__TC					3			
			BT860	M14_nonT_BT860__BC	3							
				M14_nonT_BT860__DC		5						
				M14_nonT_BT860__MC			326					
				M14_nonT_BT860__TC					4			
			NY580	M14_nonT_NY580__MC			337					
				M14_nonT_NY580__TC					4			
			NY860	M14_nonT_NY860__MC			327					
				M14_nonT_NY860__TC					8			
		M14_rTreg	NY580	M14_rTreg_NY580__MC			1					
		M14_Tncl	BT580	M14_Tncl_BT580__MC			1					
	M16	M16_aTreg	BT580	M16_aTreg_BT580__MC			4			398	0.0352	
			BT860	M16_aTreg_BT860__MC			5					
			NY580	M16_aTreg_NY580__MC			7					
			NY860	M16_aTreg_NY860__MC			7					
		M16_nonT	BT580	M16_nonT_BT580__DC		2						
				M16_nonT_BT580__MC			57					
			BT860	M16_nonT_BT860__BC	1							
				M16_nonT_BT860__DC		1						
				M16_nonT_BT860__MC			97					
				M16_nonT_BT860__TC					8			
			NY580	M16_nonT_NY580__MC			79					
				M16_nonT_NY580__TC					2			
			NY860	M16_nonT_NY860__MC			126					
		M16_T8em	BT580	M16_T8em_BT580__MC			1					
		M16_T8em	NY860	M16_T8em_NY860__MC			1					

NK_cells	NK	NK_aTreg	BT580	NK_aTreg_BT580___MC			1			1394	0.2561	1394
				NK_aTreg_BT580___TC					1			
			NY580	NK_aTreg_NY580___TC					3			
			NY860	NK_aTreg_NY860___TC					1			
		NK_nonT	BT580	NK_nonT_BT580___MC			3					
				NK_nonT_BT580___NK				216				
				NK_nonT_BT580___TC					35			
			BT860	NK_nonT_BT860___BC	2							
				NK_nonT_BT860___NK				337				
				NK_nonT_BT860___TC					86			
			NY580	NK_nonT_NY580___MC			2					
				NK_nonT_NY580___NK				160				
				NK_nonT_NY580___TC					26			
			NY860	NK_nonT_NY860___NK				216				
				NK_nonT_NY860___TC					48			
		NK_T4em	NY860	NK_T4em_NY860___TC					1			
		NK_T4naive	NY860	NK_T4naive_NY860___TC					1			
		NK_T8em	BT580	NK_T8em_BT580___NK				15				
				NK_T8em_BT580___TC					25			
			BT860	NK_T8em_BT860___NK				37				
				NK_T8em_BT860___TC					49			
			NY580	NK_T8em_NY580___NK				12				
				NK_T8em_NY580___TC					6			
			NY860	NK_T8em_NY860___NK				34				
				NK_T8em_NY860___TC					34			
		NK_Tncl	BT580	NK_Tncl_BT580___NK				2				
				NK_Tncl_BT580___TC					8			
			BT860	NK_Tncl_BT860___NK				4				
				NK_Tncl_BT860___TC					6			
			NY580	NK_Tncl_NY580___NK				1				
				NK_Tncl_NY580___TC					10			
			NY860	NK_Tncl_NY860___NK				3				
				NK_Tncl_NY860___TC					9			
T_cells	aTreg	T_aTreg	BT580	T_aTreg_BT580___MC			1			921	0.0011	8326
				T_aTreg_BT580___TC					240			
			BT860	T_aTreg_BT860___TC					243			
			NY580	T_aTreg_NY580___TC					222			
			NY860	T_aTreg_NY860___TC					215			
	nonT	T_nonT	BT580	T_nonT_BT580___NK				46		426	0.4272	
				T_nonT_BT580___TC					50			
			BT860	T_nonT_BT860___NK				51				
				T_nonT_BT860___TC					83			
			NY580	T_nonT_NY580___NK				49				
				T_nonT_NY580___TC					36			
			NY860	T_nonT_NY860___NK				36				
				T_nonT_NY860___TC					75			
	rTreg	T_rTreg	BT580	T_rTreg_BT580___MC			3			1072	0.0037	
				T_rTreg_BT580___TC					310			
			BT860	T_rTreg_BT860___MC			1					
				T_rTreg_BT860___TC					233			
			NY580	T_rTreg_NY580___TC					337			
			NY860	T_rTreg_NY860___TC					188			
	T4em	T_T4em	BT580	T_T4em_BT580___MC			1			975	0.0041	
				T_T4em_BT580___TC					329			
			BT860	T_T4em_BT860___NK				3				
				T_T4em_BT860___TC					256			
			NY580	T_T4em_NY580___TC					254			
			NY860	T_T4em_NY860___TC					132			
	T4naive	T_T4naive	BT580	T_T4naive_BT580___DC		1				1134	0.0035	
				T_T4naive_BT580___MC			1					
				T_T4naive_BT580___TC					480			
			BT860	T_T4naive_BT860___MC			1					
				T_T4naive_BT860___TC					264			
			NY580	T_T4naive_NY580___NK				1				
				T_T4naive_NY580___TC					290			
			NY860	T_T4naive_NY860___TC					96			
	T8em	T_T8em	BT580	T_T8em_BT580___MC			1			1031	0.0504	
				T_T8em_BT580___NK				11				
				T_T8em_BT580___TC					254			
			BT860	T_T8em_BT860___NK				16				
				T_T8em_BT860___TC					288			
			NY580	T_T8em_NY580___NK				11				
				T_T8em_NY580___TC					255			
			NY860	T_T8em_NY860___NK				13				
				T_T8em_NY860___TC					182			
	T8naive	T_T8naive	BT580	T_T8naive_BT580___TC					318	1336	0.0000	
			BT860	T_T8naive_BT860___TC					486			
			NY580	T_T8naive_NY580___TC					256			
			NY860	T_T8naive_NY860___TC					276			
	Tncl	T_Tncl	BT580	T_Tncl_BT580___MC			2			1431	0.0203	
				T_Tncl_BT580___NK				8				
				T_Tncl_BT580___TC					191			
			BT860	T_Tncl_BT860___MC			1					
				T_Tncl_BT860___NK				6				
				T_Tncl_BT860___TC					359			
			NY580	T_Tncl_NY580___NK				7				
				T_Tncl_NY580___TC					372			
			NY860	T_Tncl_NY860___NK				5				
				T_Tncl_NY860___TC					480			
All (predicted)					**1543**	**166**	**1696**	**1316**	**8462**	**13183**		**13183**

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
2	10x (Clean)	BC	10085	85423	√	
		M14	2612		√	
		NK	8385		√	
		CD45RA+CD2	10479		√	
		T4	11213		√	
		CD45RA+T8n	11953		√	
		T8	10209		√	
		CD45RO+T4m	10224		√	
		CD4+CD25+T	10263		√	
	GEO (of R12)	M14_d1	425	14185	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T	435		√	
					√	
					√	
					√	
					√	
					√	
					√	
					√	
		HLA-DR	48		√	
		HLA-DR_cont	2397		√	
		CD19	26		√	
		CD19_contro	1760		√	
		CD8	5662		√	
	BroadS1	Bn	1169	13183	√	
		Bm	491		√	
		DC	142		√	
		M14	1263		√	
		M16	398		√	
		NK	1394		√	
		aTreg	921		√	
		nonT	426		√	
		rTreg	1072		√	
		T4em	975		√	
		T4naive	1134		√	
		T8em	1031		√	
		T8naive	1336		√	
		Tncl	1431		√	
	BroadS2 (Clean)	BC	1884	12292		√
		DC	202			√
		pDC	68			√
		M14	1809			√
		M16	323			√
		NK	842			√
		T4	3380			√
		T8	3784			√

Accuracy:	0.898226				
Precision:	0.968421	0.5	0.86755519	0.5498008	0.954872
Recall/Ser	0.927813	0.011111	0.97701689	0.81947743	0.909687
Specificity	0.994523	0.99975	0.96870079	0.95065502	0.939938
F1_Score:	0.947682	0.021739	0.91903816	0.65808298	0.931732

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1748	0	111	15	10	1884
Dendritic_	4	3	120	2	141	270
Monocyte	31	0	2083	0	18	2132
NK_cells	4	0	9	690	139	842
T_cells	18	3	78	548	6517	7164
All	1805	6	2401	1255	6825	12292

True/ Predicted						BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	BC	pbmc1	v2	A	pbmc1_v2_A_BC__BC	239					1884	0.0722	1884
					pbmc1_v2_A_BC__MC			39					
					pbmc1_v2_A_BC__NK				6				
					pbmc1_v2_A_BC__TC					4			
				B	pbmc1_v2_B_BC__BC	351							
					pbmc1_v2_B_BC__MC			31					
					pbmc1_v2_B_BC__NK				4				
					pbmc1_v2_B_BC__TC					2			
			v3		pbmc1_v3_BC__BC	323							
					pbmc1_v3_BC__MC			19					
					pbmc1_v3_BC__NK				3				
					pbmc1_v3_BC__TC					1			
		pbmc2	v2		pbmc2_V2_BC__BC	835							
					pbmc2_V2_BC__MC			22					
					pbmc2_V2_BC__NK				2				
					pbmc2_V2_BC__TC					3			
Dendritic_cells	DC	pbmc1	v2	A	pbmc1_v2_A_DC__MC			32			202	0.9851	270
					pbmc1_v2_A_DC__TC					23			
				B	pbmc1_v2_B_DC__MC			12					
					pbmc1_v2_B_DC__TC					21			
			v3		pbmc1_v3_DC__MC			11					
					pbmc1_v3_DC__NK				2				
					pbmc1_v3_DC__TC					25			
		pbmc2	v2		pbmc2_V2_DC__DC		3						
					pbmc2_V2_DC__MC			22					
					pbmc2_V2_DC__TC					51			
	pDC	pbmc1	v2	A	pbmc1_v2_A_pDC__BC	1					68	1.0000	
					pbmc1_v2_A_pDC__MC			19					
					pbmc1_v2_A_pDC__TC					6			
				B	pbmc1_v2_B_pDC__MC			7					
					pbmc1_v2_B_pDC__TC					5			
		pbmc2	V2		pbmc2_V2_pDC__BC	3							
					pbmc2_V2_pDC__MC			17					
					pbmc2_V2_pDC__TC					10			
Monocytes	M14	pbmc1	v2	A	pbmc1_v2_A_M14__BC	19					1809	0.0216	2132
					pbmc1_v2_A_M14__MC			616					
					pbmc1_v2_A_M14__TC					5			
				B	pbmc1_v2_B_M14__BC	5							
					pbmc1_v2_B_M14__MC			372					
					pbmc1_v2_B_M14__TC					2			
			v3		pbmc1_v3_M14__MC			353					
					pbmc1_v3_M14__TC					1			
		pbmc2	V2		pbmc2_V2_M14__BC	5							
					pbmc2_V2_M14__MC			429					
					pbmc2_V2_M14__TC					2			
	M16	pbmc1	v2	A	pbmc1_v2_A_M16__BC	2					323	0.0310	
					pbmc1_v2_A_M16__MC			93					
					pbmc1_v2_A_M16__TC					7			
				B	pbmc1_v2_B_M16__MC			73					
			v3		pbmc1_v3_M16__MC			97					
					pbmc1_v3_M16__TC					1			
		pbmc2	V2		pbmc2_V2_M16__MC			50					
NK_cells	NK	pbmc1	v2	A	pbmc1_v2_A_NK__BC	1					842	0.1805	842
					pbmc1_v2_A_NK__MC			3					
					pbmc1_v2_A_NK__NK				128				
					pbmc1_v2_A_NK__TC					34			
				B	pbmc1_v2_B_NK__BC	1							
					pbmc1_v2_B_NK__MC			3					
					pbmc1_v2_B_NK__NK				189				
					pbmc1_v2_B_NK__TC					70			
			v3		pbmc1_v3_NK__BC	1							
					pbmc1_v3_NK__MC			2					
					pbmc1_v3_NK__NK				175				
					pbmc1_v3_NK__TC					16			
		pbmc2	V2		pbmc2_V2_NK__BC	1							
					pbmc2_V2_NK__MC			1					
					pbmc2_V2_NK__NK				198				
					pbmc2_V2_NK__TC					19			
T_cells	T4	pbmc1	v2	A	pbmc1_v2_A_T4__BC	2					3380	0.0195	7164
					pbmc1_v2_A_T4__MC			4					
					pbmc1_v2_A_T4__NK				5				
					pbmc1_v2_A_T4__TC					539			
				B	pbmc1_v2_B_T4__BC	1							
					pbmc1_v2_B_T4__MC			8					
					pbmc1_v2_B_T4__NK				10				
					pbmc1_v2_B_T4__TC					889			
			v3		pbmc1_v3_T4__MC			4					
					pbmc1_v3_T4__NK				15				
					pbmc1_v3_T4__TC					941			
		pbmc2	V2		pbmc2_V2_T4__BC	3							
					pbmc2_V2_T4__DC		3						
					pbmc2_V2_T4__MC			8					
					pbmc2_V2_T4__NK				3				
					pbmc2_V2_T4__TC					945			
	T8	pbmc1	v2	A	pbmc1_v2_A_T8__BC	8					3784	0.1535	
					pbmc1_v2_A_T8__MC			22					
					pbmc1_v2_A_T8__NK				174				
					pbmc1_v2_A_T8__TC					970			
				B	pbmc1_v2_B_T8__BC	1							
					pbmc1_v2_B_T8__MC			11					
					pbmc1_v2_B_T8__NK				110				
					pbmc1_v2_B_T8__TC					832			
			v3		pbmc1_v3_T8__BC	1							
					pbmc1_v3_T8__MC			10					
					pbmc1_v3_T8__NK				151				
					pbmc1_v3_T8__TC					800			
		pbmc2	V2		pbmc2_V2_T8__BC	2							
					pbmc2_V2_T8__MC			11					
					pbmc2_V2_T8__NK				80				
					pbmc2_V2_T8__TC					601			
All (predicted)						**1805**	**6**	**2401**	**1255**	**6825**	**12292**		**12292**

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
3	10x (Clean)	BC	10085	85423		√
		M14	2612			√
		NK	8385			√
		CD45RA+CD25-T4naive	10479			√
		T4	11213			√
		CD45RA+T8naive	11953			√
		T8	10209			√
		CD45RO+T4mem	10224			√
		CD4+CD25+Treg	10263			√
	GEO (of R12)	M14_d1	425	14185	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T4	435		√	
					√	
					√	
					√	
					√	
					√	
					√	
					√	
		HLA-DR	48		√	
		HLA-DR_control	2397		√	
		CD19	26		√	
		CD19_control	1760		√	
		CD8	5662		√	
	BroadS1	Bn	1169	13183	√	
		Bm	491		√	
		DC	142		√	
		M14	1263		√	
		M16	398		√	
		NK	1394		√	
		aTreg	921		√	
		nonT	426		√	
		rTreg	1072		√	
		T4em	975		√	
		T4naive	1134		√	
		T8em	1031		√	
		T8naive	1336		√	
		Tncl	1431		√	
	BroadS2 (Clean)	BC	1884	12292	√	
		DC	202		√	
		pDC	68		√	
		M14	1809		√	
		M16	323		√	
		NK	842		√	
		T4	3380		√	
		T8	3784		√	

Accuracy:	0.102080236				
Precision:	0.64285714	0	0.032306	0.992411	0.968488
Recall/Sens	7.23E-02	0	0.973201	0.14037	6.64E-02
Specificity:	0.99462423	0.99990635	0.080533	0.999883	0.993407
F1_Score:	0.12995811	0	0.062537	0.245951	0.124273

Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	729	0	9354	0	2	10085
Monocytes	6	6	2542	3	55	2612
NK_cells	1	0	7125	1177	82	8385
T_cells	398	2	59663	6	4272	64341
All	1134	8	78684	1186	4411	85423

True/ Predicted			BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	BC	021-CD19+B___BC	729					10085	0.9277	10085
		021-CD19+B___MC			9354					
		021-CD19+B___TC					2			
Monocytes	M14	003-M14___BC	6					2612	0.0268	2612
		003-M14___DC		6						
		003-M14___MC			2542					
		003-M14___NK				3				
		003-M14___TC					55			
NK_cells	NK	018-CD56+NK___BC	1					8385	0.8596	8385
		018-CD56+NK___MC			7125					
		018-CD56+NK___NK				1177				
		018-CD56+NK___TC					82			
T_cells	CD45RA+CD25-T4naive	025-CD4+CD45RA+CD25-NaiveT___BC	27					10479	0.9714	64341
		025-CD4+CD45RA+CD25-NaiveT___MC			10152					
		025-CD4+CD45RA+CD25-NaiveT___TC					300			
	T4	026-T4___BC	97					11213	0.9579	
		026-T4___DC		2						
		026-T4___MC			10642					
		026-T4___TC					472			
	CD45RA+T8naive	027-CD8+CD45RA+NaiveCytotoxicT___BC	2					11953	0.9797	
		027-CD8+CD45RA+NaiveCytotoxicT___MC			11707					
		027-CD8+CD45RA+NaiveCytotoxicT___NK				1				
		027-CD8+CD45RA+NaiveCytotoxicT___TC					243			
	T8	022-T8___BC	2					10209	0.9170	
		022-T8___MC			9357					
		022-T8___NK				3				
		022-T8___TC					847			
	CD45RO+T4mem	023-CD4+CD45RO+MemoryT___BC	7					10224	0.8990	
		023-CD4+CD45RO+MemoryT___MC			9182					
		023-CD4+CD45RO+MemoryT___NK				2				
		023-CD4+CD45RO+MemoryT___TC					1033			
	CD4+CD25+Treg	024-CD4+CD25+RegulatoryT___BC	263					10263	0.8658	
		024-CD4+CD25+RegulatoryT___MC			8623					
		024-CD4+CD25+RegulatoryT___TC					1377			
All (predicted)			**1134**	**8**	**78684**	**1186**	**4411**	**85423**		**85423**

EXP	DataSets	Subtype	SubtypeN	TotalCell	Training	Testing
4	10x (Clean)	BC	10085	85423	V	
		M14	2612		V	
		NK	8385		V	
		CD45RA+CD25-T4naive	10479		V	
		T4	11213		V	
		CD45RA+T8naive	11953		V	
		T8	10209		V	
		CD45RO+T4mem	10224		V	
		CD4+CD25+Treg	10263		V	
	GEO (of R12)	M14_d1	425	14185		V
		M14_d2	431			V
		NK	309			V
		T4	222			V
		T8	310			V
		iNKT	325			V
		MAIT	382			V
		Vd1	284			V
		Vd2	204			V
		T4	965			V
		CCR5+CD69-T4	435			V
						V
						V
						V
						V
						V
						V
						V
		HLA-DR	48			V
		HLA-DR_control	2397			V
		CD19	26			V
		CD19_control	1760			V
		CD8	5662			V
	BroadS1	Bn	1169	13183	V	
		Bm	491		V	
		DC	142		V	
		M14	1263		V	
		M16	398		V	
		NK	1394		V	
		aTreg	921		V	
		nonT	426		V	
		rTreg	1072		V	
		T4em	975		V	
		T4naive	1134		V	
		T8em	1031		V	
		T8naive	1336		V	
		Tncl	1431		V	
	BroadS2 (Clean)	BC	1884	12292	V	
		DC	202		V	
		pDC	68		V	
		M14	1809		V	
		M16	323		V	
		NK	842		V	
		T4	3380		V	
		T8	3784		V	

Accuracy:	0.80655622					
Precision:	0.81570409	0	0.853066	0.175699	0.925857	
Recall/Sensi	0.70380739	0	0.733414	0.996764	0.848219	
Specificity:	0.97709493	0.99993	0.961687	0.895863	0.889362	
F1_Score:	0.75563571	0	0.788728	0.298739	0.885339	
Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1257	1	204	79	245	1786
Monocytes	64	0	2421	465	351	3301
NK_cells	0	0	0	308	1	309
T_cells	220	0	213	901	7455	8789
All	1541	1	2838	1753	8052	14185

True/ Predicted			BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	CD19_control	GEO_GSM3258348_CD19_control___BC	1249					1760	0.2903	1786
		GEO_GSM3258348_CD19_control___MC			197					
		GEO_GSM3258348_CD19_control___NK				79				
		GEO_GSM3258348_CD19_control___TC					235			
	CD19	GEO_GSM3258346_CD19___BC	8					26	0.6923	
		GEO_GSM3258346_CD19___DC		1						
		GEO_GSM3258346_CD19___MC			7					
		GEO_GSM3258346_CD19___TC					10			
Monocytes	M14_d1	GEO_GSM2773408_M14_d1___MC			420			425	0.0118	3301
		GEO_GSM2773408_M14_d1___NK				1				
		GEO_GSM2773408_M14_d1___TC					4			
	M14_d2	GEO_GSM2773409_M14_d2___BC	3					431	0.0278	
		GEO_GSM2773409_M14_d2___MC			419					
		GEO_GSM2773409_M14_d2___NK				4				
		GEO_GSM2773409_M14_d2___TC					5			
	HLA-DR	GEO_GSM3258345_HLA-DR_BC	5					48	0.3125	
		GEO_GSM3258345_HLA-DR_MC			33					
		GEO_GSM3258345_HLA-DR_NK				3				
		GEO_GSM3258345_HLA-DR_TC					7			
	HLA-DR_control	GEO_GSM3258347_HLA-DR_control_BC	56					2397	0.3538	
		GEO_GSM3258347_HLA-DR_control_MC			1549					
		GEO_GSM3258347_HLA-DR_control_NK				457				
		GEO_GSM3258347_HLA-DR_control_TC					335			
NK_cells	NK	GEO_GSM3544603_NK___NK				308		309	0.0032	309
		GEO_GSM3544603_NK___TC					1			
T_cells	T4	GEO_20190108_GSM3544603_T4___TC					222	222	0.0000	8789
	T8	GEO_20190108_GSM3544603_T8___MC			1			310	0.0161	
		GEO_20190108_GSM3544603_T8___NK				4				
		GEO_20190108_GSM3544603_T8___TC					305			
	iNKT	GEO_20190108_GSM3544603_iNKT___NK				37		325	0.1138	
		GEO_20190108_GSM3544603_iNKT___TC					288			
	MAIT	GEO_20190108_GSM3544603_MAIT___NK				20		382	0.0524	
		GEO_20190108_GSM3544603_MAIT___TC					362			
	Vd1	GEO_20190108_GSM3544603_Vd1___MC			1			284	0.4542	
		GEO_20190108_GSM3544603_Vd1___NK				128				
		GEO_20190108_GSM3544603_Vd1___TC					155			
	Vd2	GEO_20190108_GSM3544603_Vd2___NK				44		204	0.2157	
		GEO_20190108_GSM3544603_Vd2___TC					160			
	T4	GEO_20190620_GSM3209407_T4___NK				16		965	0.0166	
		GEO_20190620_GSM3209407_T4___TC					949			
	CCR5+CD69-T4	GEO_20190620_GSM3209408_CCR5+CD69-T4___NK				9		435	0.0207	
		GEO_20190620_GSM3209408_CCR5+CD69-T4___TC					426			
	CD8	GEO_GSM3087628_T8___BC	220					5662	0.1897	
		GEO_GSM3087628_T8___MC			211					
		GEO_GSM3087628_T8___NK				643				
		GEO_GSM3087628_T8___TC					4588			
All (predicted)			**1541**	**1**	**2838**	**1753**	**8052**	**14185**	**2.7710**	**14185**

SplitConfusionMatrix-R17-clean

(R17 solely included clean data sets.)

Train: 10x(Clean)+GEO(Clean)+BroadS2(Clean)

Test: BroadS1

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
1	10x (Clean)	BC	10085	85423	√	
		M14	2612		√	
		NK	8385		√	
		CD45RA+CD25-T4na	10479		√	
		T4	11213		√	
		CD45RA+T8naive	11953		√	
		T8	10209		√	
		CD45RO+T4mem	10224		√	
		CD4+CD25+Treg	10263		√	
	GEO (Clean, R17)	M14_d1	425	4292	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T4	435		√	
	BroadS1	Bn	1169	13183		√
		Bm	491			√
		DC	142			√
		M14	1263			√
		M16	398			√
		NK	1394			√
		aTreg	921			√
		nonT	426			√
		rTreg	1072			√
		T4em	975			√
		T4naive	1134			√
		T8em	1031			√
		T8naive	1336			√
		Tncl	1431			√
	BroadS2 (Clean)	BC	1884	12292	√	
		DC	202		√	
		pDC	68		√	
		M14	1809		√	
		M16	323		√	
		NK	842		√	
		T4	3380		√	
		T8	3784		√	

Accuracy:	0.94614276					
Precision:	0.99806076	0.81437126	0.99323493	0.79407407	0.9544331	
Recall/Sensitivi	0.93012048	0.95774648	0.97230584	0.76901004	0.9735768	
Specificity:	0.99973965	0.99762288	0.9990453	0.9764187	0.9203212	
F1_Score:	0.96289367	0.8802589	0.98265896	0.78134111	0.9639099	
Predicted	B_cells	Dendritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1544	20	5	60	31	1660
Dendritic_cells	0	136	5	0	1	142
Monocytes	1	9	1615	0	36	1661
NK_cells	2	0	1	1072	319	1394
T_cells	0	2	0	218	8106	8326
All	1547	167	1626	1350	8493	13183

True/ Predicted					BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	Bn	Bn_aTreg	BT580	Bn_aTreg_BT580___BC	4					1169	0.0633	1660
			BT860	Bn_aTreg_BT860___BC	6							
			NY860	Bn_aTreg_NY860___BC	3							
		Bn_nonT	BT580	Bn_nonT_BT580___BC	237							
				Bn_nonT_BT580___DC		1						
				Bn_nonT_BT580___NK				4				
				Bn_nonT_BT580___TC					5			
			BT860	Bn_nonT_BT860___BC	519							
				Bn_nonT_BT860___DC		6						
				Bn_nonT_BT860___NK				19				
				Bn_nonT_BT860___TC					9			
			NY580	Bn_nonT_NY580___BC	153							
				Bn_nonT_NY580___DC		3						
				Bn_nonT_NY580___NK				5				
				Bn_nonT_NY580___TC					3			
			NY860	Bn_nonT_NY860___BC	171							
				Bn_nonT_NY860___DC		3						
				Bn_nonT_NY860___MC			2					
				Bn_nonT_NY860___NK				13				
				Bn_nonT_NY860___TC					1			
		Bn_T4em	BT860	Bn_T4em_BT860___BC	1							
		Bn_Tncl	BT860	Bn_Tncl_BT860___BC	1							
	Bm	Bm_aTreg	BT860	Bm_aTreg_BT860___BC	6					491	0.0855	
			NY580	Bm_aTreg_NY580___BC	1							
			NY860	Bm_aTreg_NY860___BC	2							
		Bm_nonT	BT580	Bm_nonT_BT580___BC	85							
				Bm_nonT_BT580___NK				2				
				Bm_nonT_BT580___TC					1			
			BT860	Bm_nonT_BT860___BC	207							
				Bm_nonT_BT860___DC		4						
				Bm_nonT_BT860___MC			1					
				Bm_nonT_BT860___NK				4				
				Bm_nonT_BT860___TC					7			
			NY580	Bm_nonT_NY580___BC	59							
				Bm_nonT_NY580___DC		1						
				Bm_nonT_NY580___NK				1				
				Bm_nonT_NY580___TC					2			
			NY860	Bm_nonT_NY860___BC	89							
				Bm_nonT_NY860___DC		2						
				Bm_nonT_NY860___MC			2					
				Bm_nonT_NY860___NK				12				
				Bm_nonT_NY860___TC					3			
Dendritic_cells	DC	DC_aTreg	BT860	DC_aTreg_BT860___DC		1				142	0.0423	142
			NY580	DC_aTreg_NY580___DC		1						
		DC_nonT	BT580	DC_nonT_BT580___DC		51						
				DC_nonT_BT580___MC			2					
				DC_nonT_BT580___TC					1			
			BT860	DC_nonT_BT860___DC		19						
			NY580	DC_nonT_NY580___DC		46						
			NY860	DC_nonT_NY860___DC		18						
				DC_nonT_NY860___MC			3					
Monocytes	M14	M14_aTreg	BT580	M14_aTreg_BT580___MC			1			1263	0.0269	1661
			BT860	M14_aTreg_BT860___MC			4					
			NY580	M14_aTreg_NY580___MC			2					
			NY860	M14_aTreg_NY860___MC			2					
		M14_nonT	BT580	M14_nonT_BT580___DC		2						
				M14_nonT_BT580___MC			230					
				M14_nonT_BT580___TC					6			
			BT860	M14_nonT_BT860___BC	1							
				M14_nonT_BT860___DC		4						
				M14_nonT_BT860___MC			326					
				M14_nonT_BT860___TC					7			
			NY580	M14_nonT_NY580___DC		1						
				M14_nonT_NY580___MC			335					
				M14_nonT_NY580___TC					5			
			NY860	M14_nonT_NY860___MC			327					
				M14_nonT_NY860___TC					8			
		M14_rTreg	NY580	M14_rTreg_NY580___MC			1					
		M14_Tncl	BT580	M14_Tncl_BT580___MC			1					
	M16	M16_aTreg	BT580	M16_aTreg_BT580___MC			4			398	0.0302	
			BT860	M16_aTreg_BT860___MC			5					
			NY580	M16_aTreg_NY580___MC			7					
			NY860	M16_aTreg_NY860___MC			7					
		M16_nonT	BT580	M16_nonT_BT580___DC		1						
				M16_nonT_BT580___MC			58					
			BT860	M16_nonT_BT860___DC		1						
				M16_nonT_BT860___MC			99					
				M16_nonT_BT860___TC					7			
			NY580	M16_nonT_NY580___MC			79					
				M16_nonT_NY580___TC					2			
			NY860	M16_nonT_NY860___MC			125					
				M16_nonT_NY860___TC					1			
		M16_T8em	BT580	M16_T8em_BT580___MC			1					
		M16_T8em	NY860	M16_T8em_NY860___MC			1					

NK_cells	NK	NK_aTreg	BT580	NK_aTreg_BT580__TC					2	1394	0.2310	1394
			NY580	NK_aTreg_NY580__TC					3			
			NY860	NK_aTreg_NY860__TC					1			
		NK_nonT	BT580	NK_nonT_BT580__NK				230				
				NK_nonT_BT580__TC					24			
			BT860	NK_nonT_BT860__BC	2							
				NK_nonT_BT860__NK				344				
				NK_nonT_BT860__TC					79			
			NY580	NK_nonT_NY580__MC			1					
				NK_nonT_NY580__NK				166				
				NK_nonT_NY580__TC					21			
			NY860	NK_nonT_NY860__NK				235				
				NK_nonT_NY860__TC					29			
		NK_T4em	NY860	NK_T4em_NY860__NK				1				
		NK_T4naive	NY860	NK_T4naive_NY860__TC					1			
		NK_T8em	BT580	NK_T8em_BT580__NK				11				
				NK_T8em_BT580__TC					29			
			BT860	NK_T8em_BT860__NK				26				
				NK_T8em_BT860__TC					60			
			NY580	NK_T8em_NY580__NK				11				
				NK_T8em_NY580__TC					7			
			NY860	NK_T8em_NY860__NK				38				
				NK_T8em_NY860__TC					30			
		NK_Tncl	BT580	NK_Tncl_BT580__NK				2				
				NK_Tncl_BT580__TC					8			
			BT860	NK_Tncl_BT860__NK				3				
				NK_Tncl_BT860__TC					7			
			NY580	NK_Tncl_NY580__NK				1				
				NK_Tncl_NY580__TC					10			
			NY860	NK_Tncl_NY860__NK				4				
				NK_Tncl_NY860__TC					8			
T_cells	aTreg	T_aTreg	BT580	T_aTreg_BT580__DC		1				921	0.0011	8326
				T_aTreg_BT580__TC					240			
			BT860	T_aTreg_BT860__TC					243			
			NY580	T_aTreg_NY580__TC					222			
			NY860	T_aTreg_NY860__TC					215			
	nonT	T_nonT	BT580	T_nonT_BT580__NK				42		426	0.4225	
				T_nonT_BT580__TC					54			
			BT860	T_nonT_BT860__NK				47				
				T_nonT_BT860__TC					87			
			NY580	T_nonT_NY580__NK				46				
				T_nonT_NY580__TC					39			
			NY860	T_nonT_NY860__NK				45				
				T_nonT_NY860__TC					66			
	rTreg	T_rTreg	BT580	T_rTreg_BT580__TC					313	1072	0.0000	
			BT860	T_rTreg_BT860__TC					234			
			NY580	T_rTreg_NY580__TC					337			
			NY860	T_rTreg_NY860__TC					188			
	T4em	T_T4em	BT580	T_T4em_BT580__TC					330	975	0.0000	
			BT860	T_T4em_BT860__TC					259			
			NY580	T_T4em_NY580__TC					254			
			NY860	T_T4em_NY860__TC					132			
	T4naive	T_T4naive	BT580	T_T4naive_BT580__DC		1				1134	0.0009	
				T_T4naive_BT580__TC					481			
			BT860	T_T4naive_BT860__TC					265			
			NY580	T_T4naive_NY580__TC					291			
			NY860	T_T4naive_NY860__TC					96			
	T8em	T_T8em	BT580	T_T8em_BT580__NK				6		1031	0.0281	
				T_T8em_BT580__TC					260			
			BT860	T_T8em_BT860__NK				9				
				T_T8em_BT860__TC					295			
			NY580	T_T8em_NY580__NK				8				
				T_T8em_NY580__TC					258			
			NY860	T_T8em_NY860__NK				6				
				T_T8em_NY860__TC					189			
	T8naive	T_T8naive	BT580	T_T8naive_BT580__TC					318	1336	0.0000	
			BT860	T_T8naive_BT860__TC					486			
			NY580	T_T8naive_NY580__TC					256			
			NY860	T_T8naive_NY860__TC					276			
	Tncl	T_Tncl	BT580	T_Tncl_BT580__TC					201	1431	0.0063	
			BT860	T_Tncl_BT860__NK				1				
				T_Tncl_BT860__TC					365			
			NY580	T_Tncl_NY580__NK				4				
				T_Tncl_NY580__TC					375			
			NY860	T_Tncl_NY860__NK				4				
				T_Tncl_NY860__TC					481			
All (predicted)					**1547**	**167**	**1626**	**1350**	**8493**	**13183**		**13183**

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
2	10x (Clean)	BC	10085	85423	√	
		M14	2612		√	
		NK	8385		√	
		CD45RA+CD2	10479		√	
		T4	11213		√	
		CD45RA+T8n	11953		√	
		T8	10209		√	
		CD45RO+T4m	10224		√	
		CD4+CD25+T	10263		√	
	GEO (Clean, R17)	M14_d1	425	4292	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T	435		√	
	BroadS1	Bn	1169	13183	√	
		Bm	491		√	
		DC	142		√	
		M14	1263		√	
		M16	398		√	
		NK	1394		√	
		aTreg	921		√	
		nonT	426		√	
		rTreg	1072		√	
		T4em	975		√	
		T4naive	1134		√	
		T8em	1031		√	
		T8naive	1336		√	
		Tncl	1431		√	
	BroadS2 (Clean)	BC	1884	12292		√
		DC	202			√
		pDC	68			√
		M14	1809			√
		M16	323			√
		NK	842			√
		T4	3380			√
		T8	3784			√

Accuracy:	0.917345				
Precision:	0.930983	0	0.92384682	0.55555556	0.988289
Recall/Ser	0.995223	0	0.99577861	0.9263658	0.907035
Specificity	0.986645	0.999917	0.98277559	0.94550218	0.984984
F1_Score:	0.962032	0	0.95846501	0.69456812	0.94592

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	1875	0	6	0	3	1884
Dendritic_	103	0	152	0	15	270
Monocyte	6	0	2123	0	3	2132
NK_cells	6	0	0	780	56	842
T_cells	24	1	17	624	6498	7164
All	2014	1	2298	1404	6575	12292

True/ Predicted						BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
B_cells	BC	pbmc1	v2	A	pbmc1_v2_A_BC___BC	286					1884	0.0048	1884
					pbmc1_v2_A_BC___MC			1					
					pbmc1_v2_A_BC___TC					1			
				B	pbmc1_v2_B_BC___BC	385							
					pbmc1_v2_B_BC___MC			3					
			v3		pbmc1_v3_BC___BC	345							
					pbmc1_v3_BC___MC			1					
		pbmc2	v2		pbmc2_V2_BC___BC	859							
					pbmc2_V2_BC___MC			1					
					pbmc2_V2_BC___TC					2			
Dendritic_cells	DC	pbmc1	v2	A	pbmc1_v2_A_DC___BC	6					202	1.0000	270
					pbmc1_v2_A_DC___MC			47					
					pbmc1_v2_A_DC___TC					2			
				B	pbmc1_v2_B_DC___BC	2							
					pbmc1_v2_B_DC___MC			29					
					pbmc1_v2_B_DC___TC					2			
			v3		pbmc1_v3_DC___BC	8							
					pbmc1_v3_DC___MC			24					
					pbmc1_v3_DC___TC					6			
		pbmc2	v2		pbmc2_V2_DC___BC	25							
					pbmc2_V2_DC___MC			50					
					pbmc2_V2_DC___TC					1			
	pDC	pbmc1	v2	A	pbmc1_v2_A_pDC___BC	25					68	1.0000	
					pbmc1_v2_A_pDC___MC			1					
				B	pbmc1_v2_B_pDC___BC	9							
					pbmc1_v2_B_pDC___TC					3			
		pbmc2	V2		pbmc2_V2_pDC___BC	28							
					pbmc2_V2_pDC___MC			1					
					pbmc2_V2_pDC___TC					1			
Monocytes	M14	pbmc1	v2	A	pbmc1_v2_A_M14___BC	2					1809	0.0039	2132
					pbmc1_v2_A_M14___MC			636					
					pbmc1_v2_A_M14___TC					2			
				B	pbmc1_v2_B_M14___BC	1							
					pbmc1_v2_B_M14___MC			378					
			v3		pbmc1_v3_M14___MC			354					
		pbmc2	V2		pbmc2_V2_M14___BC	2							
					pbmc2_V2_M14___MC			434					
	M16	pbmc1	v2	A	pbmc1_v2_A_M16___BC	1					323	0.0062	
					pbmc1_v2_A_M16___MC			100					
					pbmc1_v2_A_M16___TC					1			
				B	pbmc1_v2_B_M16___MC			73					
			v3		pbmc1_v3_M16___MC			98					
		pbmc2	V2		pbmc2_V2_M16___MC			50					
NK_cells	NK	pbmc1	v2	A	pbmc1_v2_A_NK___BC	2					842	0.0736	842
					pbmc1_v2_A_NK___NK				156				
					pbmc1_v2_A_NK___TC					8			
				B	pbmc1_v2_B_NK___NK				220				
					pbmc1_v2_B_NK___TC					43			
			v3		pbmc1_v3_NK___BC	3							
					pbmc1_v3_NK___NK				187				
					pbmc1_v3_NK___TC					4			
		pbmc2	V2		pbmc2_V2_NK___BC	1							
					pbmc2_V2_NK___NK				217				
					pbmc2_V2_NK___TC					1			
T_cells	T4	pbmc1	v2	A	pbmc1_v2_A_T4___BC	4					3380	0.0210	7164
					pbmc1_v2_A_T4___MC			2					
					pbmc1_v2_A_T4___NK				12				
					pbmc1_v2_A_T4___TC					532			
				B	pbmc1_v2_B_T4___MC			4					
					pbmc1_v2_B_T4___NK				12				
					pbmc1_v2_B_T4___TC					892			
			v3		pbmc1_v3_T4___BC	2							
					pbmc1_v3_T4___MC			1					
					pbmc1_v3_T4___NK				18				
					pbmc1_v3_T4___TC					939			
		pbmc2	V2		pbmc2_V2_T4___BC	3							
					pbmc2_V2_T4___DC		1						
					pbmc2_V2_T4___MC			4					
					pbmc2_V2_T4___NK				8				
					pbmc2_V2_T4___TC					946			
	T8	pbmc1	v2	A	pbmc1_v2_A_T8___BC	12					3784	0.1572	
					pbmc1_v2_A_T8___MC			6					
					pbmc1_v2_A_T8___NK				193				
					pbmc1_v2_A_T8___TC					963			
				B	pbmc1_v2_B_T8___BC	1							
					pbmc1_v2_B_T8___NK				110				
					pbmc1_v2_B_T8___TC					843			
			v3		pbmc1_v3_T8___NK				152				
					pbmc1_v3_T8___TC					810			
		pbmc2	V2		pbmc2_V2_T8___BC	2							
					pbmc2_V2_T8___NK				119				
					pbmc2_V2_T8___TC					573			
All (predicted)						**2014**	**1**	**2298**	**1404**	**6575**	**12292**		**12292**

EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
3	10x (Clean)	BC	10085	85423		√
		M14	2612			√
		NK	8385			√
		CD45RA+CD25-T4naive	10479			√
		T4	11213			√
		CD45RA+T8naive	11953			√
		T8	10209			√
		CD45RO+T4mem	10224			√
		CD4+CD25+Treg	10263			√
	GEO (Clean, R17)	M14_d1	425	4292	√	
		M14_d2	431		√	
		NK	309		√	
		T4	222		√	
		T8	310		√	
		iNKT	325		√	
		MAIT	382		√	
		Vd1	284		√	
		Vd2	204		√	
		T4	965		√	
		CCR5+CD69-T4	435		√	
	BroadS1	Bn	1169	13183	√	
		Bm	491		√	
		DC	142		√	
		M14	1263		√	
		M16	398		√	
		NK	1394		√	
		aTreg	921		√	
		nonT	426		√	
		rTreg	1072		√	
		T4em	975		√	
		T4naive	1134		√	
		T8em	1031		√	
		T8naive	1336		√	
		Tncl	1431		√	
	BroadS2 (Clean)	BC	1884	12292	√	
		DC	202		√	
		pDC	68		√	
		M14	1809		√	
		M16	323		√	
		NK	842		√	
		T4	3380		√	
		T8	3784		√	

Accuracy:	0.98292				
Precision:	0.976932	0	0.84933	0.985147	0.992145
Recall/Sens	9.62E-01	0	0.897779	0.925462	9.97E-01
Specificity:	0.99696	0.997787	0.994977	0.998481	0.975904
F1_Score:	0.969218	0	0.872883	0.954372	0.994667

Predicted	B_cells	ritic_cells	Monocytes	NK_cells	T_cells	All
B_cells	9698	28	356	1	2	10085
Monocytes	202	19	2345	3	43	2612
NK_cells	0	135	27	7760	463	8385
T_cells	27	7	33	113	64161	64341
All	9927	189	2761	7877	64669	85423

True/ Predicted			BC	DC	MC	NK	TC	SubtypeN	SubtypeE	All (true)
B_cells	BC	021-CD19+B___BC	9698					10085	0.0384	10085
		021-CD19+B___DC		28						
		021-CD19+B___MC			356					
		021-CD19+B___NK				1				
		021-CD19+B___TC					2			
Monocytes	M14	003-M14___BC	202					2612	0.1022	2612
		003-M14___DC		19						
		003-M14___MC			2345					
		003-M14___NK				3				
		003-M14___TC					43			
NK_cells	NK	018-CD56+NK___DC		135				8385	0.0745	8385
		018-CD56+NK___MC			27					
		018-CD56+NK___NK				7760				
		018-CD56+NK___TC					463			
T_cells	CD45RA+CD25-T4naive	025-CD4+CD45RA+CD25-NaiveT___BC	7					10479	0.0042	64341
		025-CD4+CD45RA+CD25-NaiveT___DC		6						
		025-CD4+CD45RA+CD25-NaiveT___MC			15					
		025-CD4+CD45RA+CD25-NaiveT___NK				16				
		025-CD4+CD45RA+CD25-NaiveT___TC					10435			
	T4	026-T4___BC	9					11213	0.0021	
		026-T4___DC		1						
		026-T4___MC			5					
		026-T4___NK				9				
		026-T4___TC					11189			
	CD45RA+T8naive	027-CD8+CD45RA+NaiveCytotoxicT___BC	5					11953	0.0009	
		027-CD8+CD45RA+NaiveCytotoxicT___MC			4					
		027-CD8+CD45RA+NaiveCytotoxicT___NK				2				
		027-CD8+CD45RA+NaiveCytotoxicT___TC					11942			
	T8	022-T8___MC			7			10209	0.0079	
		022-T8___NK				74				
		022-T8___TC					10128			
	CD45RO+T4mem	023-CD4+CD45RO+MemoryT___BC	1					10224	0.0003	
		023-CD4+CD45RO+MemoryT___MC			2					
		023-CD4+CD45RO+MemoryT___TC					10221			
	CD4+CD25+Treg	024-CD4+CD25+RegulatoryT___BC	5					10263	0.0017	
		024-CD4+CD25+RegulatoryT___NK				12				
		024-CD4+CD25+RegulatoryT___TC					10246			
All (predicted)			**9927**	**189**	**2761**	**7877**	**64669**	**85423**		**85423**

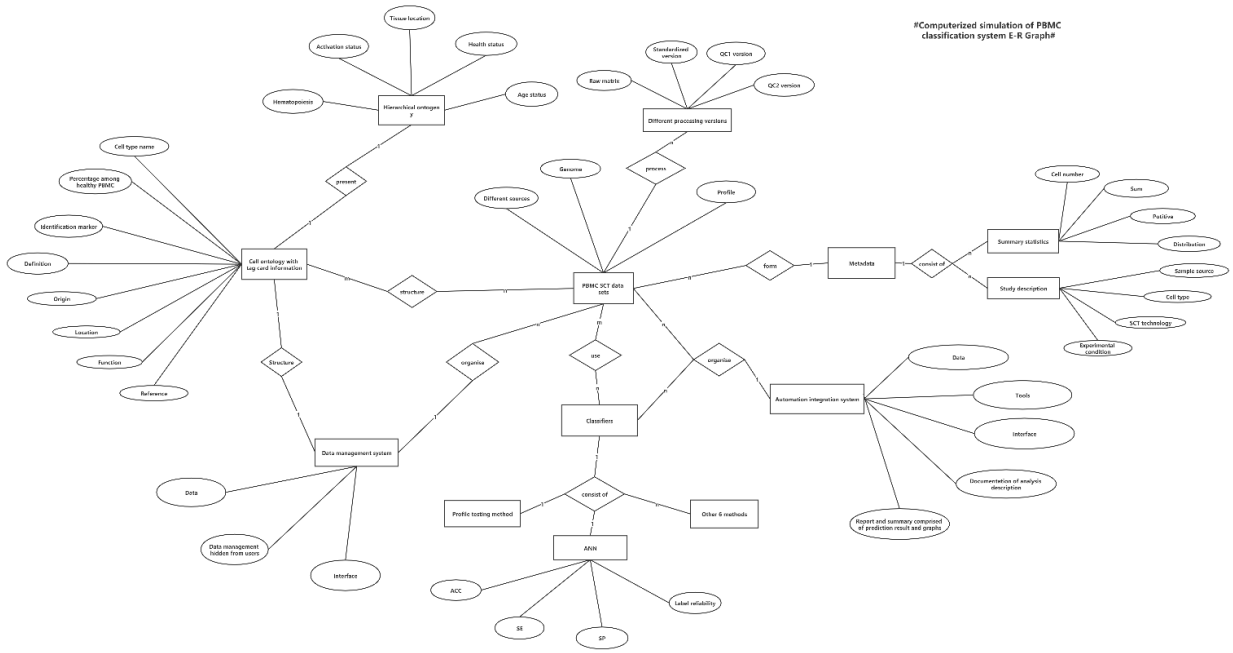
EXP	DataSets	Subtype	SubtypeN	TotalCellN	Training	Testing
4	10x (Clean)	BC	10085	85423	V	
		M14	2612		V	
		NK	8385		V	
		CD45RA+CD25-T4naive	10479		V	
		T4	11213		V	
		CD45RA+T8naive	11953		V	
		T8	10209		V	
		CD45RO+T4mem	10224		V	
		CD4+CD25+Treg	10263		V	
	GEO (Clean, R17)	M14_d1	425	4292		V
		M14_d2	431			V
		NK	309			V
		T4	222			V
		T8	310			V
		iNKT	325			V
		MAIT	382			V
		Vd1	284			V
		Vd2	204			V
		T4	965			V
		CCR5+CD69-T4	435			V
	BroadS1	Bn	1169	13183	V	
		Bm	491		V	
		DC	142		V	
		M14	1263		V	
		M16	398		V	
		NK	1394		V	
		aTreg	921		V	
		nonT	426		V	
		rTreg	1072		V	
		T4em	975		V	
		T4naive	1134		V	
		T8em	1031		V	
		T8naive	1336		V	
		Tncl	1431		V	
	BroadS2 (Clean)	BC	1884	12292	V	
		DC	202		V	
		pDC	68		V	
		M14	1809		V	
		M16	323		V	
		NK	842		V	
		T4	3380		V	
		T8	3784		V	

Accuracy:	0.93522833			
Precision:	0	0.997622	0.539405	0.99652416
Recall/Sensi	0	0.98014	0.996764	0.91685321
Specificity:	0.99930103	0.999418	0.933969	0.99141631
F1_Score:	0	0.988804	0.7	0.95502998

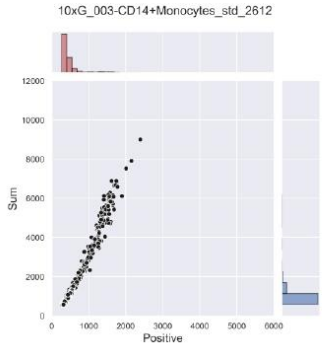
Predicted	B_cells	Monocytes	NK_cells	T_cells	All
Monocytes	3	839	5	9	856
NK_cells	0	0	308	1	309
T_cells	0	2	258	2867	3127
All	3	841	571	2877	4292

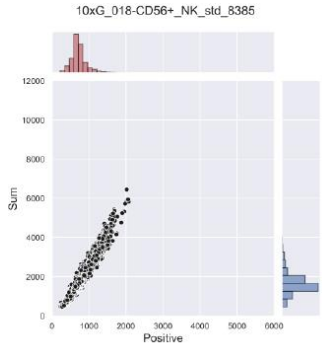
True/ Predicted			BC	DC	MC	NK	TC	SubtypeN	SubtypeER	All (true)
Monocytes	M14_d1	GEO_GSM2773408_M14_d1___MC			420			425	0.0118	856
		GEO_GSM2773408_M14_d1___NK				1				
		GEO_GSM2773408_M14_d1___TC					4			
	M14_d2	GEO_GSM2773409_M14_d2___BC	3					431	0.0278	
		GEO_GSM2773409_M14_d2___MC			419					
		GEO_GSM2773409_M14_d2___NK				4				
		GEO_GSM2773409_M14_d2___TC					5			
NK_cells	NK	GEO_20190108_GSM3544603_NK___NK				308		309	0.0032	309
		GEO_20190108_GSM3544603_NK___TC					1			
T_cells	T4	GEO_20190108_GSM3544603_T4___TC					222	222	0.0000	3127
	T8	GEO_20190108_GSM3544603_T8___MC			1			310	0.0161	
		GEO_20190108_GSM3544603_T8___NK				4				
		GEO_20190108_GSM3544603_T8___TC					305			
	iNKT	GEO_20190108_GSM3544603_iNKT___NK				37		325	0.1138	
		GEO_20190108_GSM3544603_iNKT___TC					288			
	MAIT	GEO_20190108_GSM3544603_MAIT___NK				20		382	0.0524	
		GEO_20190108_GSM3544603_MAIT___TC					362			
	Vd1	GEO_20190108_GSM3544603_Vd1___MC			1			284	0.4542	
		GEO_20190108_GSM3544603_Vd1___NK				128				
		GEO_20190108_GSM3544603_Vd1___TC					155			
	Vd2	GEO_20190108_GSM3544603_Vd2___NK				44		204	0.2157	
		GEO_20190108_GSM3544603_Vd2___TC					160			
	T4	GEO_20190620_GSM3209407_T4___NK				16		965	0.0166	
		GEO_20190620_GSM3209407_T4___TC					949			
	CCR5+CD69-T4	GEO_20190620_GSM3209408_CCR5+CD69-T4___NK				9		435	0.0207	
		GEO_20190620_GSM3209408_CCR5+CD69-T4___TC					426			
All (predicted)			**3**	**0**	**841**	**571**	**2877**	**4292**		**4292**

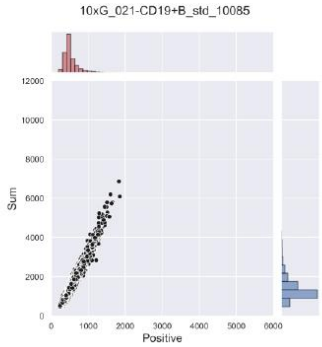
Appendix 9 E-R Graph of This Project

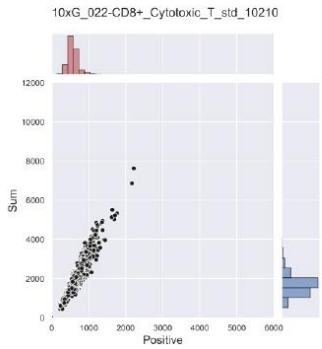


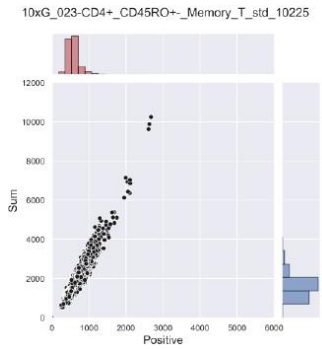
Appendix 10 Visualization of SCT Data Distribution

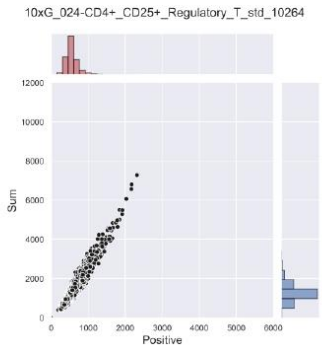


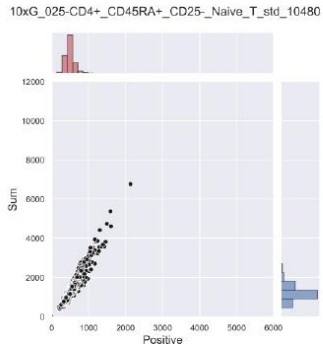


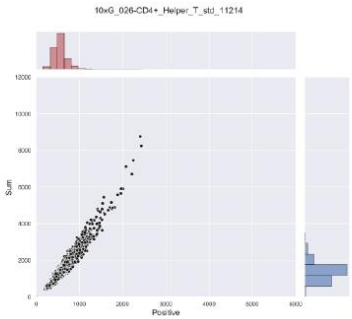


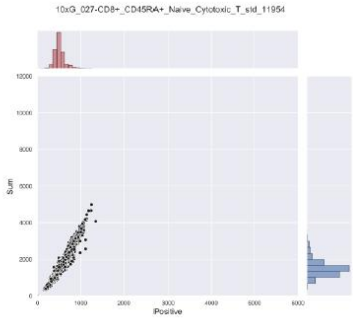


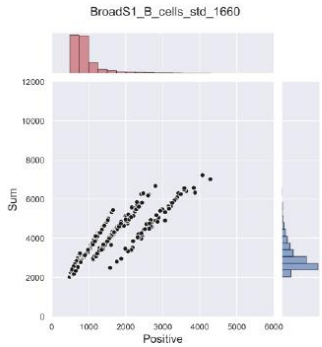


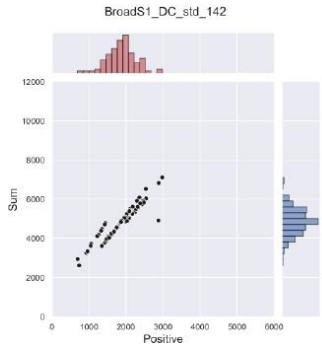


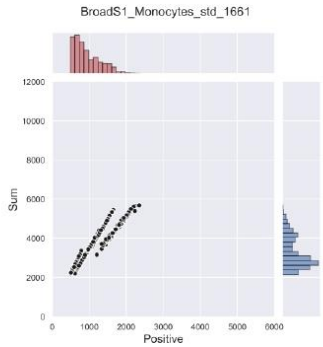


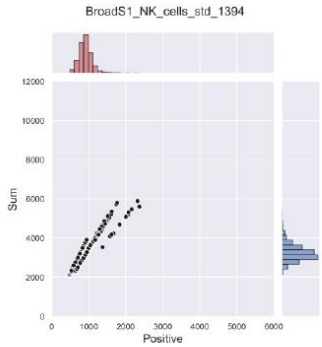


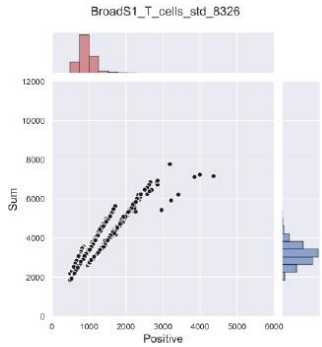


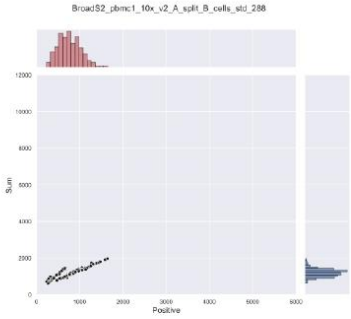

BroadS1_B_cells_std_1660
Sum
Positive

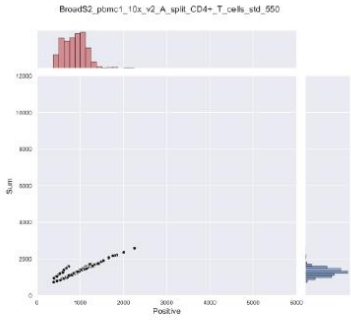

BroadS1_DC_std_142
Sum
Positive

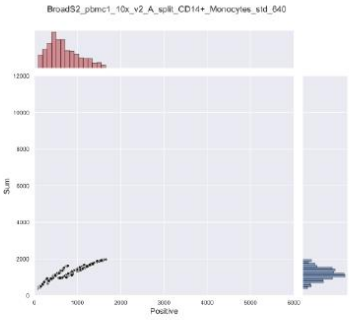

BroadS1_Monocytes_std_1661
Sum
Positive

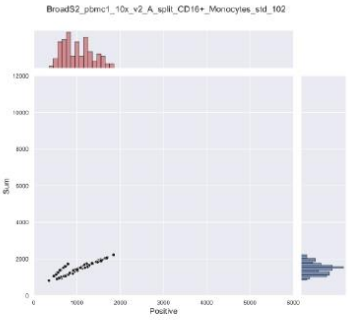

BroadS1_NK_cells_std_1394
Sum
Positive

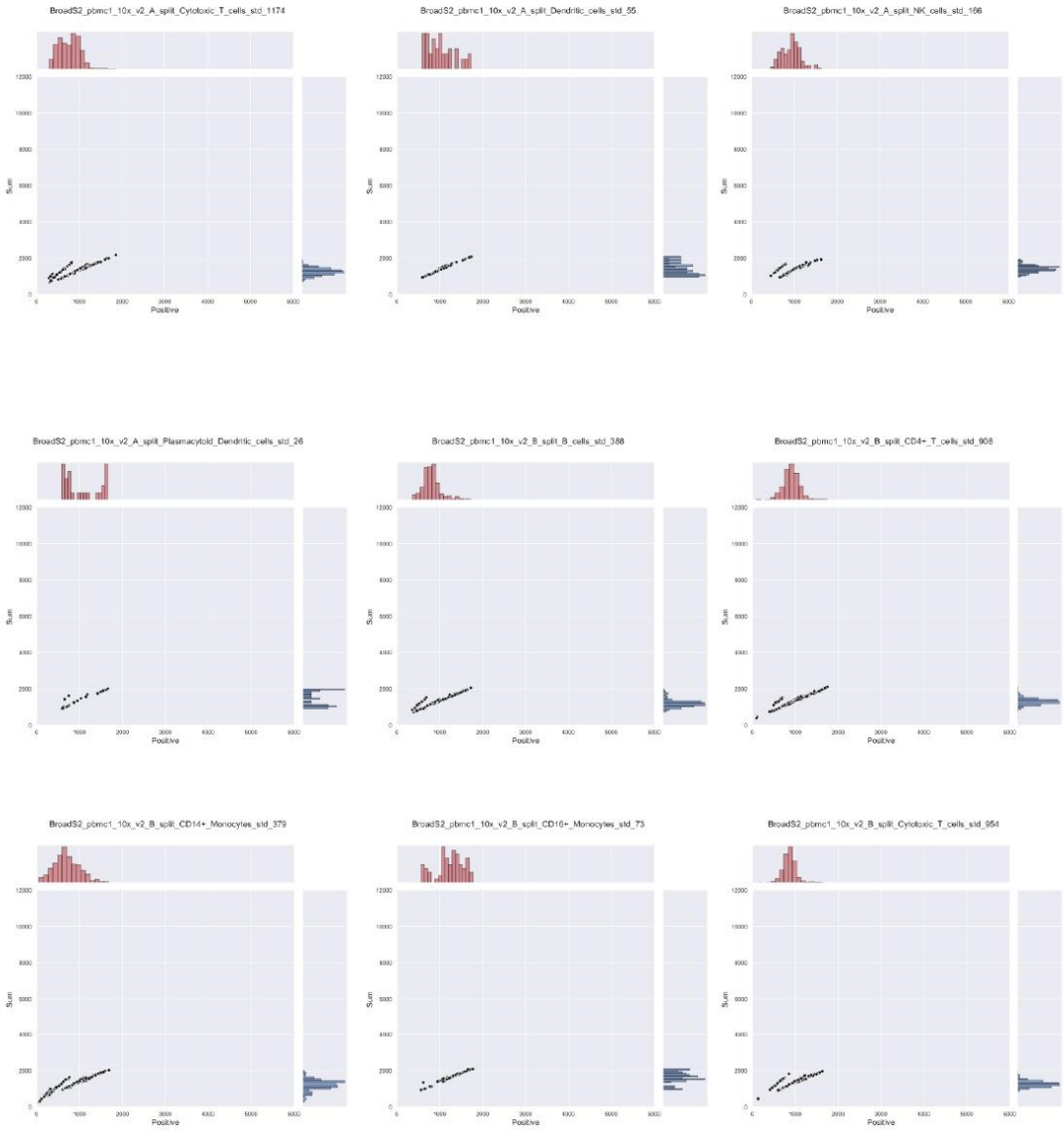

BroadS1_T_cells_std_8326
Sum
Positive

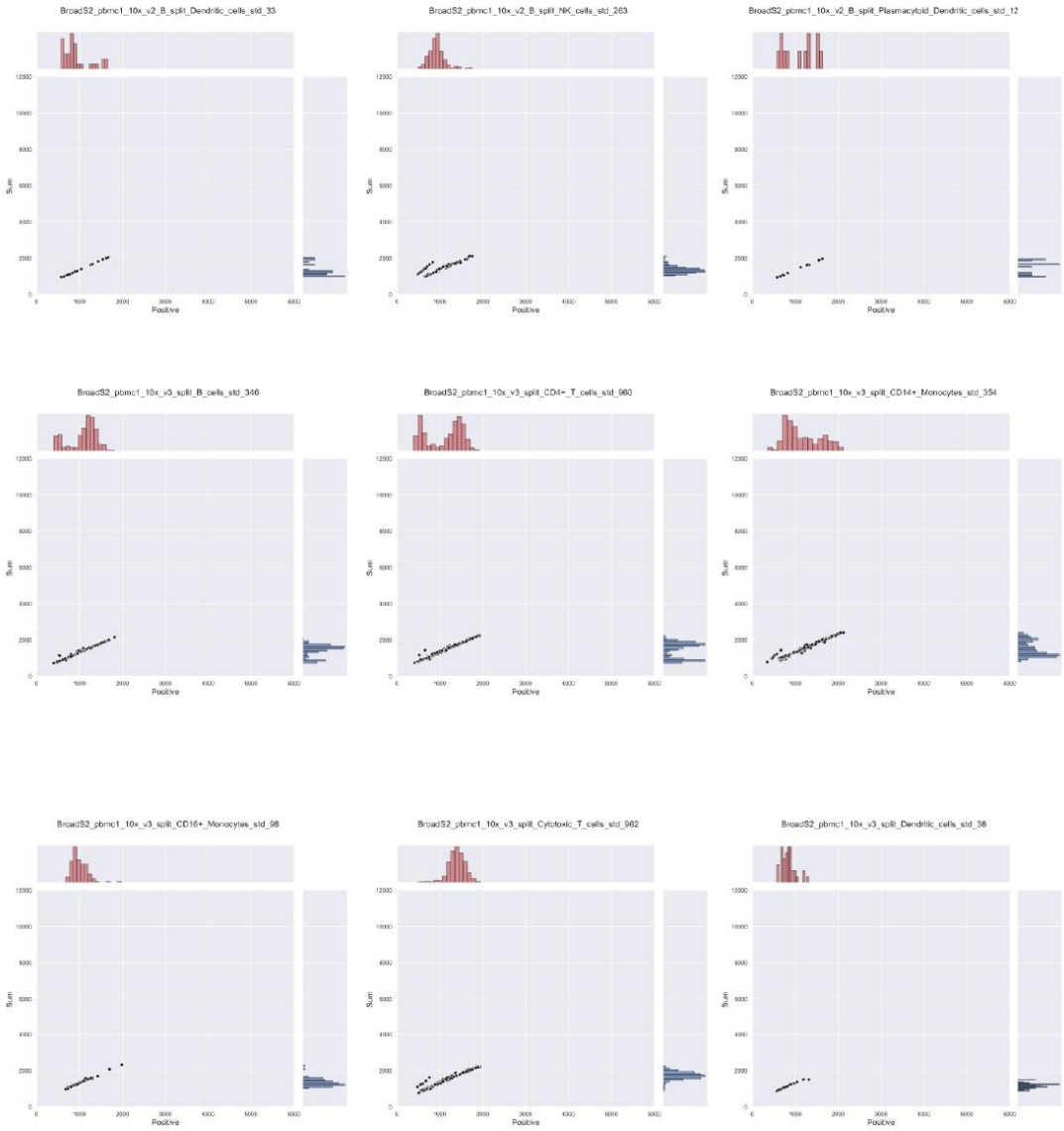

BroadS2_pbmc1_10x_v2_A_split_B_cells_std_288

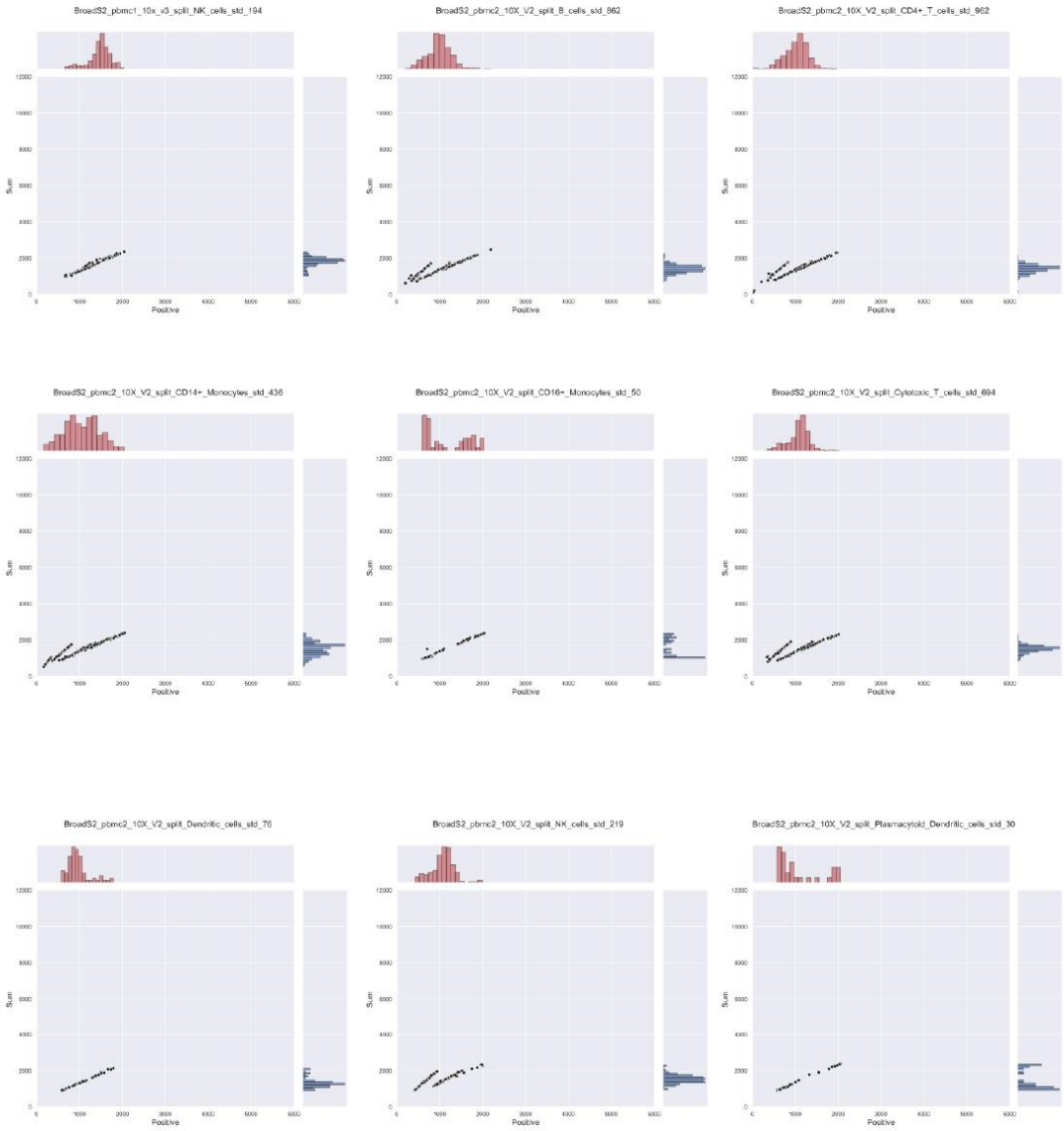

BroadS2_pbmc1_10x_v2_A_split_CD4+_T_cells_std_550

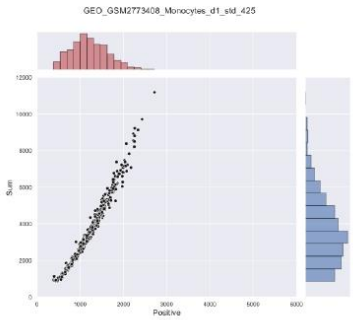

BroadS2_pbmc1_10x_v2_A_split_CD14+_Monocytes_std_640

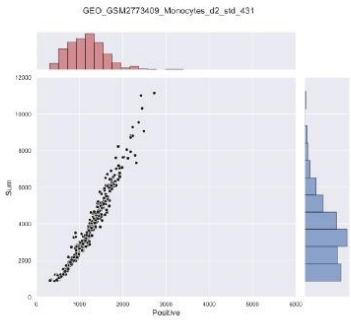

BroadS2_pbmc1_10x_v2_A_split_CD16+_Monocytes_std_102

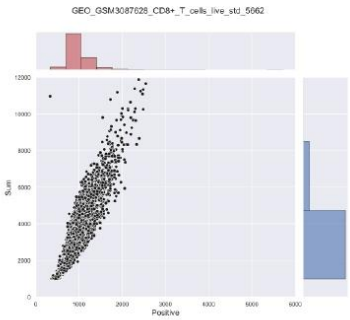

BroadS2_pbmc1_10x_v2_A_split_Cytotoxic_T_cells_std_1174
BroadS2_pbmc1_10x_v2_A_split_Dendritic_cells_std_55
BroadS2_pbmc1_10x_v2_A_split_NK_cells_std_166
BroadS2_pbmc1_10x_v2_A_split_Plasmacytoid_Dendritic_cells_std_26
BroadS2_pbmc1_10x_v2_B_split_B_cells_std_388
BroadS2_pbmc1_10x_v2_B_split_CD4+_T_cells_std_908
BroadS2_pbmc1_10x_v2_B_split_CD14+_Monocytes_std_379
BroadS2_pbmc1_10x_v2_B_split_CD16+_Monocytes_std_73
BroadS2_pbmc1_10x_v2_B_split_Cytotoxic_T_cells_std_954
Sum
Positive

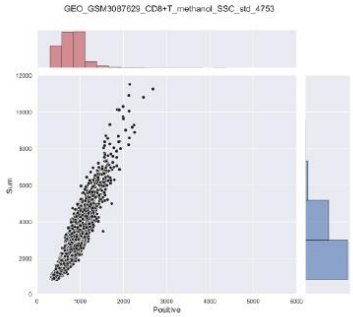

BroadS2_pbmc1_10x_v2_B_split_Dendritic_cells_std_33
BroadS2_pbmc1_10x_v2_B_split_NK_cells_std_283
BroadS2_pbmc1_10x_v2_B_split_Plasmacytoid_Dendritic_cells_std_12
BroadS2_pbmc1_10x_v3_split_B_cells_std_346
BroadS2_pbmc1_10x_v3_split_CD4+_T_cells_std_960
BroadS2_pbmc1_10x_v3_split_CD14+_Monocytes_std_354
BroadS2_pbmc1_10x_v3_split_CD16+_Monocytes_std_98
BroadS2_pbmc1_10x_v3_split_Cytotoxic_T_cells_std_962
BroadS2_pbmc1_10x_v3_split_Dendritic_cells_std_38
Sum
Positive

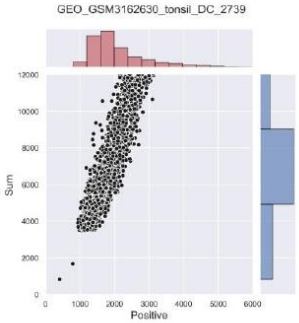

BroadS2_pbmc1_10x_v3_split_NK_cells_std_194
BroadS2_pbmc2_10X_V2_split_B_cells_std_862
BroadS2_pbmc2_10X_V2_split_CD4+_T_cells_std_962
BroadS2_pbmc2_10X_V2_split_CD14+_Monocytes_std_436
BroadS2_pbmc2_10X_V2_split_CD16+_Monocytes_std_50
BroadS2_pbmc2_10X_V2_split_Cytotoxic_T_cells_std_694
BroadS2_pbmc2_10X_V2_split_Dendritic_cells_std_76
BroadS2_pbmc2_10X_V2_split_NK_cells_std_219
BroadS2_pbmc2_10X_V2_split_Plasmacytoid_Dendritic_cells_std_30
Sum
Positive

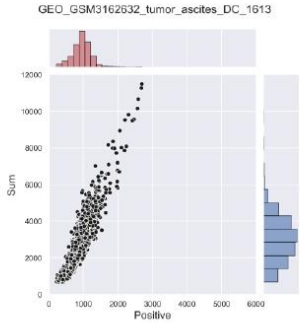

GEO_GSM2773408_Monocytes_d1_std_425
Sum
Positive

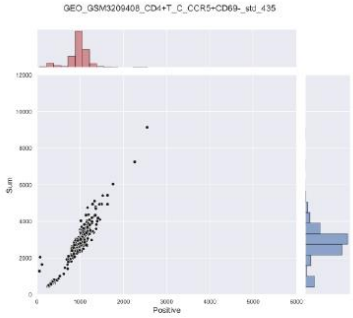

GEO_GSM2773409_Monocytes_d2_std_431
Sum
Positive

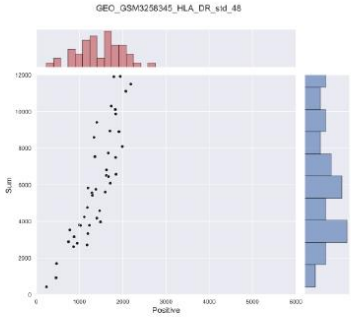

GEO_GSM3087628_CD8+_T_cells_live_std_5662
Sum
Positive

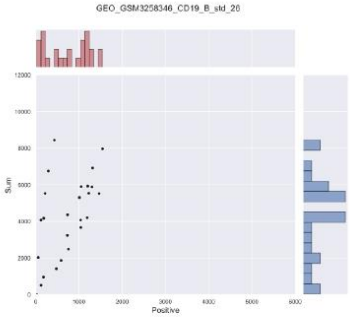

GEO_GSM3087629_CD8+T_methanol_SSC_std_4753
Sum
Positive

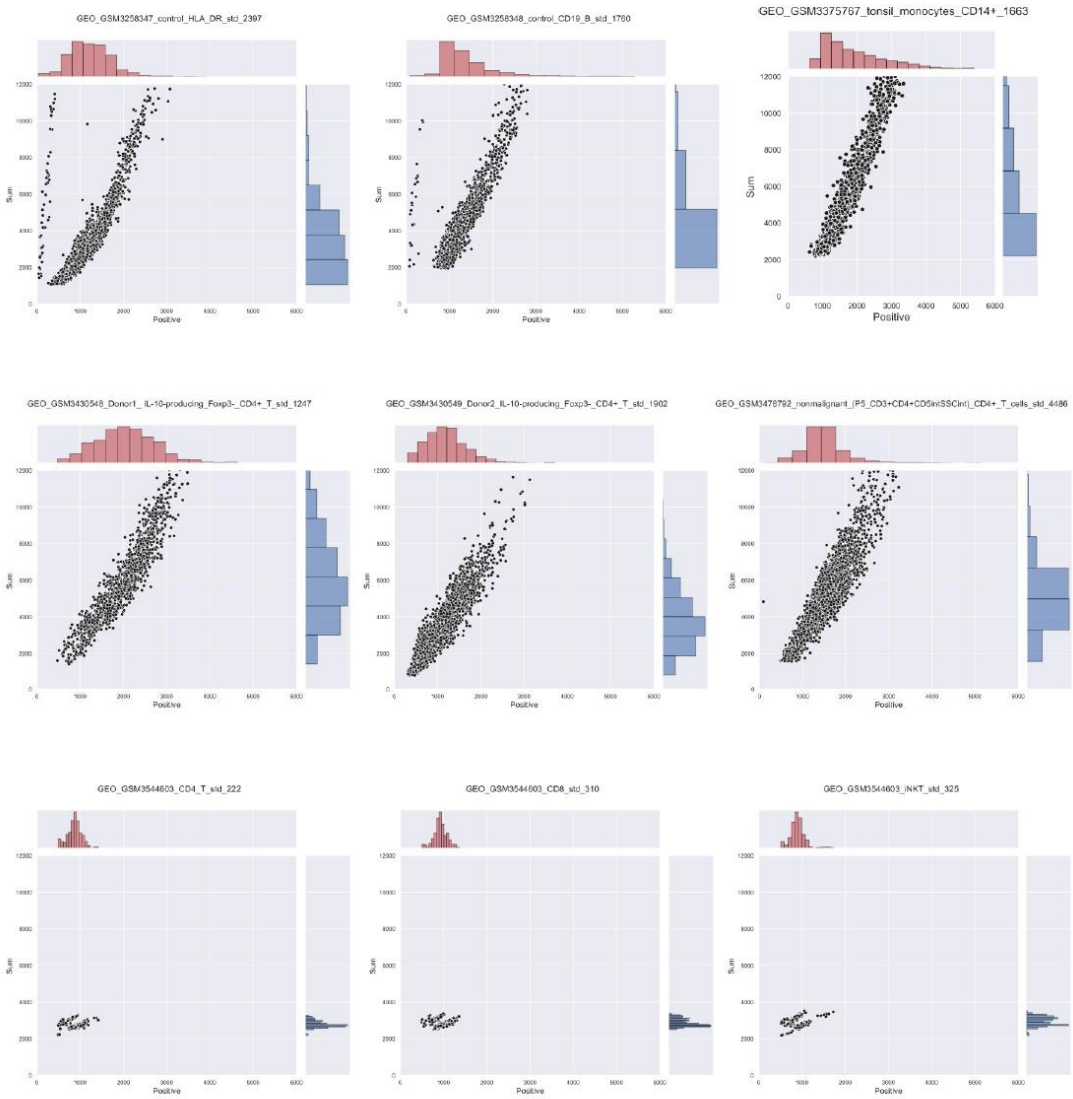

GEO_GSM3162630_tonsil_DC_2739
Sum
Positive

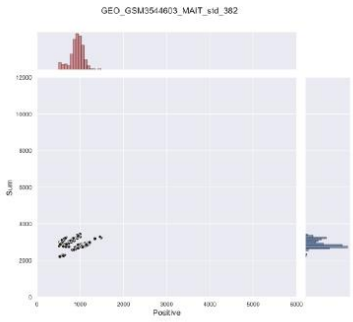

GEO_GSM3162632_tumor_ascites_DC_1613
Sum
Positive

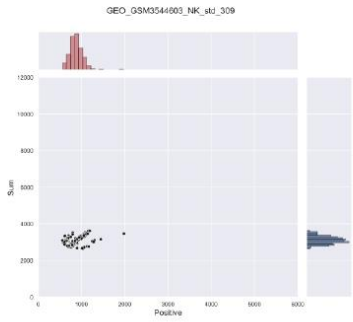

GEO_GSM3209408_CD4+T_C_CCR5+CD69-_std_435
Sum
Positive

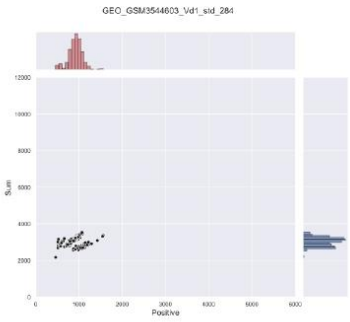

GEO_GSM3258345_HLA_DR_std_48
Sum
Positive

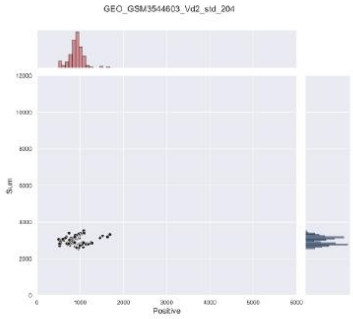

GEO_GSM3258346_CD19_B_std_20
Sum
Positive

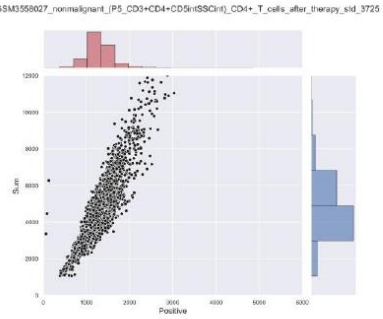

GEO_GSM3258347_control_HLA_DR_std_2397
GEO_GSM3258348_control_CD19_B_std_1760
GEO_GSM3375767_tonsil_monocytes_CD14+_1663
GEO_GSM3430548_Donor1_ IL-10-producing_Foxp3-_CD4+_T_std_1247
GEO_GSM3430549_Donor2_IL-10-producing_Foxp3-_CD4+_T_std_1902
GEO_GSM3478792_nonmalignant_(P5_CD3+CD4+CD5intSSCint)_CD4+_T_cells_std_4486
GEO_GSM3544603_CD4_T_std_222
GEO_GSM3544603_CD8_std_310
GEO_GSM3544603_iNKT_std_325
Sum
Positive


GEO_GSM3544603_MAIT_sid_382
Sum
Positive


GEO_GSM3544603_NK_sid_309
Sum
Positive


GEO_GSM3544603_Vd1_sid_284
Sum
Positive


GEO_GSM3544603_Vd2_sid_204
Sum
Positive


GEO_GSM3558027_nonmalignant_(P5_CD3+CD4+CD5intSSCint)_CD4+_T_cells_after_therapy_sid_3725
Sum
Positive

Appendix 11 Posters During This Project

University of Nottingham
UK | CHINA | MALAYSIA

University of Nottingham Ningbo China

Classification of cells using single cell transcriptomics data and machine learning

Jiahui ZHONG, Vladimir BRUSIC

BACKGROUND & MOTIVATION

In recent years, single cell transcriptomics (SCT) becomes much popular research method instead of bulk sequencing technology. It can detect heterogeneous genetic information which is not obtained by mixed sample multicellular sequencing. This leads the whole field of genetics into a new dimension. Downstream analysis to single cell experimental data with supervised machine learning can build classification of cells, detect rare subtype of blood cells, refine the ontology of immune cells and conduct diseases diagnosis and health prediction. ***This project focus on classification of cells using SCT data and machine learning.*** The objective of this project is to develop and implement an applied system that can use the updating single cell database, perform supervised machine learning, and validate the performance of the system on the follow-up new data sets. The project includes the following steps: data collection, feature extraction, model structure building and training, classification system training, testing and validation, and comparative analysis. ***This system will be used to determine the cell type and the sorting of one single cell.*** The goal is to provide a proof of concept that a Machine Learning system based on supervised learning can accurately classify cells from mixed samples.

AIMS

(1) Collect and standardize a large and diverse data set to serve as input for classification system.

(2) Develop and implement a supervised machine learning method for classification of multi-class samples of SCT measurements.

(3) Generalize the learning model across multiple independent studies.

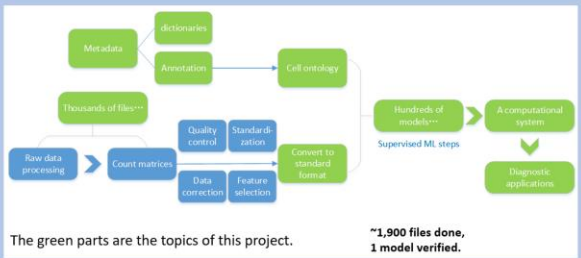


Figure 1: scRNA-seq analysis workflow using supervised machine learning methods.

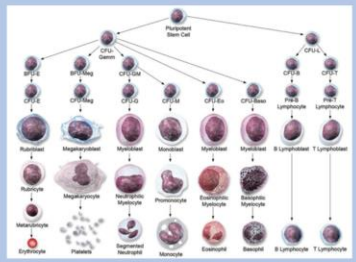


Figure 2: Heterogeneity of blood cells. Barreto et al., *Journal of Pharmacy Practice* 27(5):440-446, (2014)

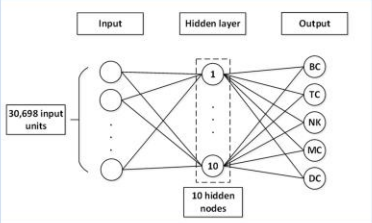


Figure 3: Methods – ANN Architecture.

- Cycle 7: ANN trained using all 10xS + GEOS data sets.
- High accuracy relative to previous cycles, using an independent test set.
- Accuracy = 89.4%

TABLE V. CYCLE 7 CONFUSION MATRIX

Experimental \ Predicted	PBMC BC	TA+TO DC	PBMC MC	PBMC NK	PBMC TC	SUM
PBMC BC	1,624	7	102	2	16	1,751
PBMC DC	0	69	72	0	1	142
PBMC MC	120	143	1,324	2	79	1,668
PBMC NK	23	11	4	1,110	246	1,394
PBMC TC	55	10	58	464	7,757	8,344
SUM	1,822	240	1,560	1,578	8,099	13,299

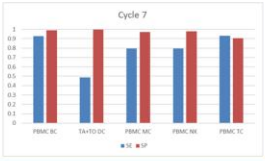


Figure 4: Prospective validation result (Cycle 7) of the first case study – PBMCs subtype classification and prediction.

Methodology

Step 1 - SCT data collection and organization - Done

Step 2 - Concept demonstration of big data analysis with artificial neural network (ANN) – (Done)

Step 3 - Optimization of model architecture and algorithm – (In progress)

Step 4 - Supervised machine learning with cell ontology annotation

Step 5 - Generalize the classification method to diversified SCT data

CHALLENGES

Reported SCT data are very noisy, lack standardized format, are Big Data.

We have ***biological, technical (experimental), and data processing noise.***

Contacts: Jiahui ZHONG jiahui.zhong@Nottingham.edu.cn
Vladimir BRUSIC vladimir.brusic@Nottingham.edu.cn

University of Nottingham
UK | CHINA | MALAYSIA

University of Nottingham Ningbo China

Classification of cells using single cell transcriptomics data and machine learning

Jiahui ZHONG, Vladimir BRUSIC

INTRODUCTION

Single cell transcriptomics (SCT) detects gene expression profiles from individual cells. Determining gene expression from mixed sample by bulk sequencing loses information about heterogeneity of gene expression between cell types and subtypes [1]. Combining cell sorting, SCT sequencing and machine learning (ML) enables classification of cells and even the discovery of novel cell subtypes [2]. Current ML methods focus on unsupervised clustering for assigning class labels to single cells from their transcriptomics profiles. This approach may be suitable for the analysis of individual data sets in combination with various annotations techniques, but unsupervised clustering has limited accuracy and generalizes poorly across different studies [3].

We applied artificial neural networks (ANN) a supervised ML technique to demonstrate improved accuracy and generalization of classification of PBMC. In our previous work we demonstrated the accuracy of five-class classification of human peripheral blood mononuclear cells (PBMC) to be approximately 90% [4]. In the current study we examined properties of data sets, analyzed feature extraction options, and performed incremental learning by increasing the number of training data sets. Here we report new data pre-processing method and improved ANN models for classification of PBMC and discuss the implications of supervised ML to solving challenges of single cell transcriptomic analysis.

MATERIALS AND METHODS

In this study, we collected, selected, cleaned and standardized 27 10x SCT data sets of healthy fresh blood sample that represent expression profiles of PBMC subtypes. ANN incremental learning model was trained to classify 5 main cell types of PBMC and demonstrate the initial concept of computerize supervised learning SCT cell classification system. The ANN incremental learning model should perform good classification accuracy and have well robustness across various SCT data derived from different studies.

1. **Data organization** – Collect, clean and standardize a large, sparse, noisy and diverse **10x SCT PBMC data sets** to serve as input for classification system.
2. **Cell ontology** on hierarchical class annotation of PBMC data sets.
3. **Concept demonstration** of multi-class SCT data sets classification with implementing supervised machine learning method **artificial neural network (ANN).**
4. **Incremental learning** on optimization of model architecture and algorithm.
5. **Generalize** the performance of classification method across multiple diversified independent SCT data sets.

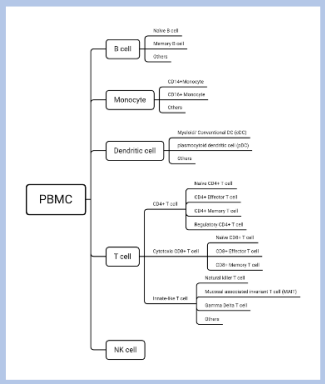


Figure 1 Heterogeneity of human peripheral blood mononuclear cells (PBMC) base on actual data sets used in this study.

Cell type/ Sources	10x Gen	GEO DB	BroadS1	Total
B cells	1	1	1	3
Dendritic cells	0	0	1	1
Monocytes	1	2	1	4
NK cells	1	1	1	3
T cells	6	9	1	16
Total	**9**	**13**	**5**	**27**

Table 1 The number of data sets mainly used in this project study.

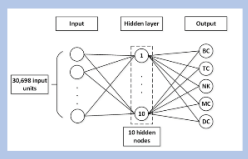


Figure 2 The current-in-use ANN model architecture.

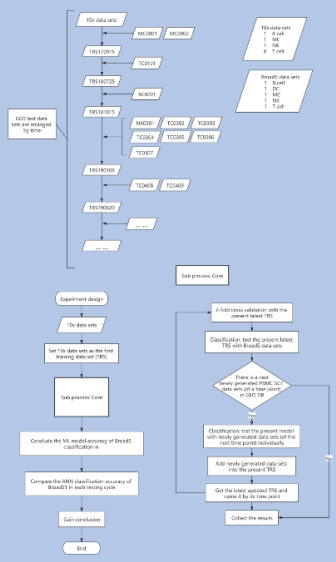


Figure 3 Incremental learning structure and experimental design for computerized cell classification using SCT data and supervised machine learning methods.

RESULTS

ANN classification to 10x SCT data sets with multi source data sets has been done with incremental learning based on 10x demonstration data, BroadS1 data and GEO DB data. The performance results of incremental learning with all data sets has been shown as below.

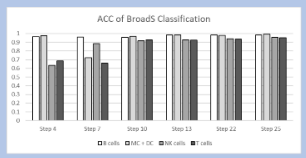


Figure 4 ANN performance on cell type classification of the incremental learning experiment across different cycle steps.

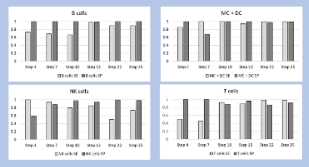


Figure 5 ANN predication performance on each cell type in the incremental learning experiment.

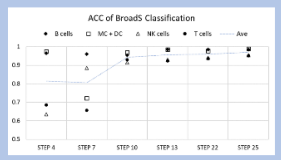


Figure 6 The line chart of ANN performance with incremental learning across different cycle steps.

CONCLUSION & DISCUSSION

1. Compared to the previous research [4], the incremental learning experiment design has eliminated the tonsil dendritic cell data set, while adding two other GEO data sets in.
   The ANN prediction accuracy has been increased from 89.4% to 92.9%.
2. The classification performance **has been gradually improved** in the process with incremental data sets collecting in, the ANN has demonstrated overall good robustness in the final learning step.
3. This incremental learning study has been done with only one same prediction model. The performance needs to be generally validated and demonstrated **on other different models.**
4. **New data set** needs to be implemented to show **the generalization of ANN model classification** to diverse multi-sources 10x SCT data.
5. **Biological noise and technical noise** can bias the recognition performance of ANN model.
6. The impact of **data quality control**, **data distribution** (low-end & high-end, cell number) and **study experimental condition** on ANN incremental learning need to be figured out for each individual data set.

FOLLOWING WORK

- Generalize the robustness of ANN performance with 10 different random models.
- Optimize the architecture of ANN model to reach better prediction performance and robustness to multi-source diverse data sets.
- The impact of quality control on data distribution and removing zeros in gene expression on ANN model need to be figured out.
- Cell subtype class confirmation and recognition based on ANN and data distribution/structure analysis.
- Refine cell ontology.
- Discover and define new cell type.

References

1. Kulkarni A, et al., *Curr Opin Biotech* 2019;58:129.
2. Villani AC, et al. Science 2017; 356(6335):eaah4573.
3. Chen L, et al. 2020. *Genes* 2020;11(7):792.
4. Shaikh RA, et al. Proc. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 2207-2213). IEEE.

Contacts: Jiahui ZHONG jiahui.zhong@nottingham.edu.cn
Vladimir BRUSIC vladimir.brusic@nottingham.edu.cn

University of Nottingham UK | CHINA | MALAYSIA

FoSE Faculty of Science and Engineering University of Nottingham Ningbo China

SSNI Jiahui Zhong 20120196

Artificial Neural Networks for Classification of Single Cell Gene Expression

Jiahui Zhong, Prof. Vladimir Brusic

Single Cell Transcriptomics Group, SSNI, UNNC

INTRODUCTION

Single cell transcriptomics (SCT) can detect heterogeneous genetic information which is not obtained by mixed sample multicellular sequencing. This leads the whole field of genetics into a new dimension. Computerized cell classification with SCT data sets using supervised machine learning can bring specific labeled learning and classification procedure to each individual single cell gene count expression profile, where is improved to unsupervised machine learning clustering and biological manual cell sorting FACS.

Cell classification with SCT data and artificial neural network (ANN) is aim to achieve to build classification of single cells, detect rare subtype of blood cells, refine the ontology of immune cells and conduct diseases diagnosis and health prediction. This research has demonstrated the classification system with five main cell types of peripheral blood mononuclear cells (PBMC).

MATERIALS AND METHODS

I. **Data organization** – Collect, clean and standardize a large, sparse, noisy and diverse **10x SCT PBMC data sets** to serve as input for classification system. (Done)

II. **Cell ontology** on hierarchical class annotation of PBMC data sets. (Done)

III. **Concept demonstration** of multi-class SCT data sets classification with implementing supervised machine learning method **artificial neural network (ANN)**. (Done)

IV. **Incremental learning** on optimization of model architecture and algorithm. (In progress)

V. **Generalize** the performance of classification method across multiple diversified independent SCT data sets. (In progress)

Source/ Cell Type (number of data sets)	B cell	Monocyte	NK cell	T cell	Dendritic cell	Total
10x Gen Demo Data	10,085 (1)	2,612 (1)	8,385 (1)	64,347 (6)	0	85,429 (9)
GEO Data	1,760 (1)	856 (2)	309 (1)	8,789 (9)	0	11,714 (13)
BroadS1	1,660 (1)	1,661 (1)	1,394 (1)	8,326 (1)	142 (1)	13,183 (5)
BroadS2	1,884 (4)	2,132 (8)	842 (4)	7,164 (8)	270 (8)	12,292 (32)
Total	15,389 (7)	7,261 (12)	10,930 (7)	88,626 (24)	412 (9)	122,618 (59)

Table 1. Detailed components of 59 SCT data sets involved in this study, sorted under five classes (five main cell types of PBMC). The number of data sets have been shown in the brackets. These data sets have been arranged into training sets and testing sets for each cycle in learning process.

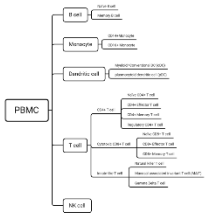


Figure 1. The ontology illustration of PBMC, that demonstrates the taxonomy and the heterogeneity of blood cells in nature. The subtypes of five main cell types of PBMC have been clarified based on collected 10x SCT data sets.

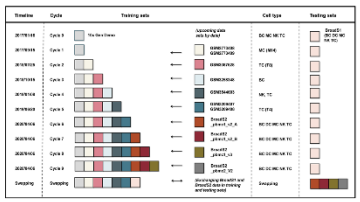


Figure 2. The detailed components of training and testing data sets in each periodic cycle in incremental learning. In each next coming up periodic cycle, the data sets with next publication date have been added into the previous training data sets to form the new current training data set. Newly formed current training data set has been implemented into the next periodic ANN training and testing cycle. The cycles and data sets have been ordered and set up to their publication dates (to simulate the actual situation in real life). At last cycle, BroadS1 and BroadS2 data sets have been swapped to observe the ANN performance on generalization properties.

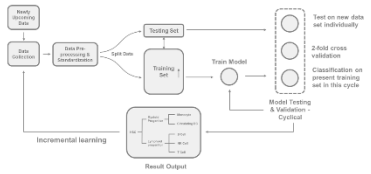


Figure 3. The overall single cell classification system for this study. SCT data has been collected, standardized, and split into training set and testing set. Single cell classification model has been trained with multi-source cumulative data sets. In each cycle of model testing and validation, testing on individual new data set, 2-fold cross validation, and classification on present training set have been demonstrated. The outputs have shown the five-class classification result. During keeping collecting and feeding with newly upcoming data sets, the cyclical system has been driven by incremental learning method.

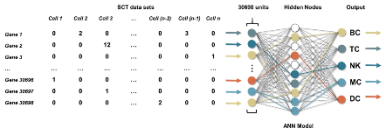


Figure 4. The artificial neural network model architecture has been employed in this study, that comprises of one input layer, one hidden layer with 10 hidden notes, and one output layer. The input layer has 30,698 input units, that refers to 30,698 gene features in the data sets. The output layer has five output units that refers to five cell classes of PBMC. The model has been trained with incremental training sets, while tested with well-annotated high-quality testing set. The activation function ReLU has been used in this model, parameters in detail have been documented in text below. The model has recognized different transcriptional expression patterns across different cell types, training with well-labeled PBMC SCT data sets.

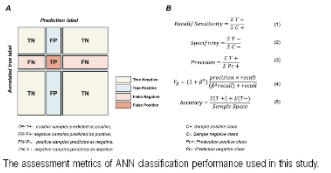


Figure 5. The assessment metrics of ANN classification performance used in this study. Confusion matrix is a visual model evaluation method, that consists of four situations to the result – TN, TP, FN, and FP. Recall/Sensitivity, specificity, precision, F1 score and accuracy have been used to measure the capability of ANN classifier.

RESULTS

ANN classification to 10x SCT data sets with multi source data sets has been done with incremental learning based on 10x demonstration data, BroadS1 data, BroadS2 data and GEO DB data. The performance results of incremental learning with all data sets has been shown.

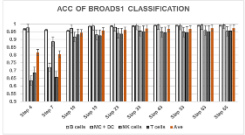


Figure 6. ANN performance on cell type classification of incremental learning process across different cycle steps. The classification result has demonstrated the classification accuracy of the final step in each training cycle tested with BroadS1 data sets. Within the increasing number of the training data sets, the performance of ANN model classification accuracy on all cell types has shown the trend of steady increasing.

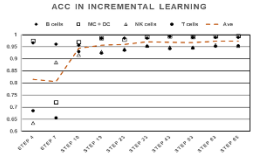


Figure 7. The overall average classification accuracy of ANN on each cell type across individual cyclical testing steps during incremental learning. During incremental learning process, the overall accuracy across B cells, Monocytes + DC, NK cells and T cells have been increased from 0.815 to 0.973, from Cycle 0 (Step 4, Figure 7.) to the final Cycle 9 (Step 65, Figure 7.).

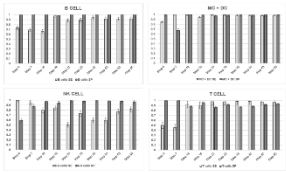


Figure 8. The result of sensitivity and specificity of B cells, Monocytes + Dendritic cells, NK cells and T cells in each training and testing cycle.

Annotated/ Predicted	B cell	Dendritic cell	Monocyte	NK cell	T cell	SUM
B cell	**1,527**	22	22	63	26	1,660
Dendritic cell	0	134	7	0	1	142
Monocyte	2	32	**1,611**	0	16	1,661
NK cell	2	0	2	**1,136**	254	1,394
T cell	2	2	1	297	**8,024**	8,326
SUM	1,533	190	1,643	1,496	8,321	**13,183**

Table 2. The confusion matrix of the final training and testing step 65. In step 65, 1,527 B cells have been correctly predicted as B cells (91.99%), 22 B cells predicted as DC, 22 B cells predicted as monocytes, 63 B cells predicted as NK cells, 26 B cells predicted as T cells. The classification accuracy of DC is 94.37%, of monocytes is 96.99%, of NK cells is 81.49%, of T cells is 96.37%. The overall accuracy across all cell types of PBMC is 94.30%.

Assessment/ Cell type	B cell	Dendritic cell	Monocyte	NK cell	T cell
Precision	0.99609611	0.70526316	0.98052343	0.75935829	0.96430717
Recall/Sensitivity	0.91987952	0.94366197	0.96989765	0.81492109	0.96372808
Specificity	0.9994793	0.99570585	0.9972227	0.96946306	0.93885114
F1_Score	0.95646727	0.80722892	0.9751816	0.78615917	0.96401754
Accuracy	**0.943032694**				

Table 3. The assessment metrics of the final training and testing step 65 in Cycle 9. The ANN classification model performance on precision, recall/sensitivity, specificity, F1_Score, accuracy of B cell class, Dendritic cell class, Monocyte class, NK cell class, and T cell class have been listed in this table.

CONCLUSION & DISCUSSION

I. The efficiency of **ANN multi-classification prototype** has been proved on blood cell PBMC classification with diverse high-dimensional and sparse 10x SCT data sets.

II. **Supervised machine learning** has demonstrated good universality and robustness for independent data sets with multiple data sources.

III. The classification performance **has been gradually improved** in the incremental learning process. The classification system has achieved high accuracy, good robustness, and good generalization ability with incremental learning method.

IV. The ANN model has demonstrated overall good robustness and **high accuracy (94.30%)** on PBMC five-class classification in incremental learning process.

V. The model performance needs to be generally validated and demonstrated **on other different models.**

VI. The effect of biological variants and technical variants brought from **data set processing protocols** on SCT expression profiles needs to be figured out in further study. It can bias the recognition performance of ANN model.

VII. The optimization of ANN architecture should be considered based on hidden layers and hidden nodes.

PUBLICATIONS

Zhong J, Shaikh RA, Wu H, Lin X, Cao Z, Chitkushev LT, Zhang G, Keskin DB and Brusic V. Classification of PBMC cell types using scRNAseq, ANN, and incremental learning. In IEEE Int. Conf. Bioinformatics Biomed. 2020, 1351-1355.

Lin X, Zhong J, Lyu M, Lin S, Keskin DB, Zhang G, Brusic V and Chitkushev LT. Artificial Neural Network System for Cell Classification using Single Cell RNA Expression. In IEEE Int. Conf. Bioinformatics Biomed. 2020, 1253-1257.

Shaikh RA, Zhong J, Lyu M, Lin S, Keskin DB, Zhang G, Chitkushev LT and Brusic V. Classification of Five Cell Types from PBMC Samples using Single Cell Transcriptomics and Artificial Neural Networks. In IEEE Int. Conf. Bioinformatics Biomed. 2019, 2207-2213.

ACKNOWLEDGEMENTS

We much thank Guanglan Zhang and Lou T. Chitkushev, from Boston University, Derin B. Keskin, from Harvard Medical School, Dana-Farber Cancer Institute, and Zhiwei Cao, from Tongji University, for their help and support for this research.

CONTACT INFORMATION

Prof. Vladimir BRUSIC Vladimir.Brusic@nottingham.edu.cn

Jiahui ZHONG Jiahui.Zhong@nottingham.edu.cn

199 Taikang East Road, Yinzhou, Ningbo, China 315100

The 3rd Annual FoSE PGR Research Showcase

New Library, UNNC 21st May 2021

Peripheral Blood Mononuclear Cell Classification using Single-cell RNA-seq Data and Artificial Neural Networks

Jiahui ZHONG, Prof. Vladimir BRUSIC, Prof. Heshan DU, Prof. Huan JIN
Single Cell Transcriptomics Group, SSNI, UNNC

BACKGROUND

Single-cell RNA-seq (SCT) can detect heterogeneous genetic information for individual cell. Single cell classification has met challenges in lack of classification method and reference data sets.

Our research [1-4] has proved the concept that single cell classification can be done with SCT data and supervised machine learning (ML) method artificial neural networks (ANN) with high accuracy, where is improved to unsupervised ML clustering and biological manual cell sorting FACS/MACS. We have made reference data sets and demonstrated a classification system using five main cell types of peripheral blood mononuclear cells (PBMC). Here, this exhibited study builds on an extension of our previous research [1]. This study aims to analyze the misclassification, model vulnerability, and data quality in PBMC single-cell classification, with four super sets external validation experiments.

MATERIALS AND METHODS

- **DATA**

This study involves SCT data sets of four different data sources: 10xDemo data, GEO, BroadS1, and BroadS2.

SOURCE/CELLTYPE (number of data sets)	B CELL	DENDRITIC CELL	MONOCYTE	NK CELL	T CELL	TOTAL
10x DEMO	10,085 (1)	0	2,612 (1)	8,385 (1)	64,341 (6)	85,423 (9)
GEO	1,760 (1)	0	856 (2)	309 (1)	8,789 (9)	11,714 (13)
BROADS1	1,660 (1)	142 (1)	1,661 (1)	1,394 (1)	8,326 (1)	13,183 (5)
BROADS2	1,884 (4)	270 (7)	2,132 (8)	842 (4)	7,164 (8)	12,292 (31)
TOTAL	15,389 (7)	412 (8)	7,261 (12)	10,930 (7)	88,620 (24)	122,612 (58)

Table 1. Detailed components of 58 SCT data sets involved in this study, sorted under five classes (BC, DC, MC, NK, and TC). The number of data sets have been shown in the brackets. These data sets have been arranged into training sets and testing sets for external validation experiments.

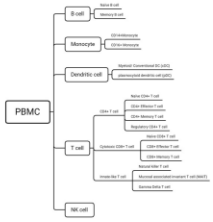


Figure 1. The ontology illustration of PBMC, that demonstrates the taxonomy and the heterogeneity of blood cells in nature. The subtypes of five main cell types of PBMC have been diarified based on collected 10x SCT data sets.

- **ANN MODEL AND ASSESSMENT METRICS**

The architecture of ANN model used in this study is as described in [1].

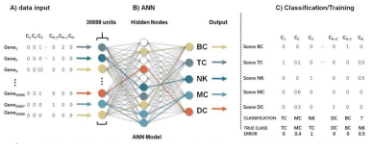


Figure 2. The ANN model comprises of one input layer, one hidden layer with 10 hidden notes, and one output layer. The input layer has 30,698 input units, that refers to 30,698 gene features in the data sets. The output layer has five output units that refers to five cell classes of PBMC. The activation function ReLU has been used in this model. The model is trained with well-labeled PBMC SCT data sets.

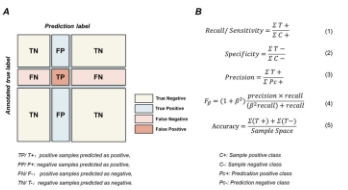
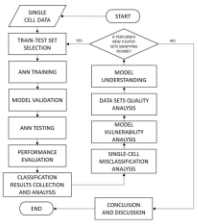


Figure 3. The assessment metrics of ANN model used in this study. Confusion matrix is a visual model evaluation method, that consists of four situations to the result – TN, TP, FN, and FP. Recall/Sensitivity, specificity, precision, F1 score and accuracy have been used to measure the capability of ANN classifier.

- **STUDY DESIGN**

In this study, four sources' data sets have been split into train set and test set for super sets swapping train-test validation. There has been three rounds of 4-sets swapping: the original one as described in previous study [1], the second round removed seven empty cells, and the third round further removed two GEO sets 'CD19' and 'TC-5662', for analysis of misclassification, data quality, and model vulnerability.



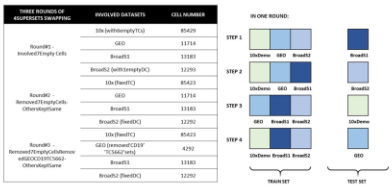


Figure 4. The workflow of study design and the detail of train-test set in three rounds of 4-super-sets-swapping experiments.

RESULTS

- **COMPARATIVE ANALYSIS**

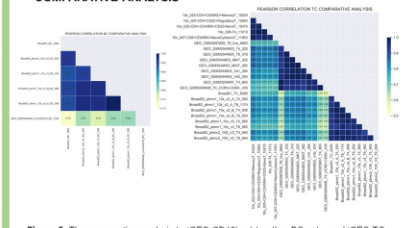


Figure 5. The comparative analysis to 'GEO-CD19' set to other BC sets, and 'GEO-TC-5662' set to other TC sets of four data sources, using Pearson correlation heatmap.

- The comparative analysis of 'GEO-CD19', 'GEO-TC-5662' to other BC, TC data sets has been done. The results has shown 'GEO-CD19' set and 'GEO-TC-5662' sets have different gene expression profiles from other BC and TC SCT data sets.
- The average of correlation coefficient value of 'GEO-CD19' to other BC sets was 0.77 (compared to 0.96 of others). The TC set 'GEO-TC-5662' has shown weak association with other involved TC sets (correlation coefficient 0.56~0.67).

- **OVERALL ACCURACY**

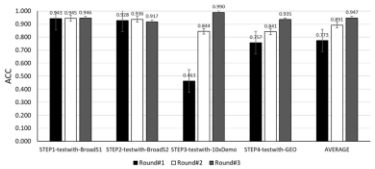


Figure 6. The overall ACC of each swapping train-test step of three rounds.

- The classification model has been vulnerable for 5.87% when train set involving misleading data sets (R#2), for 18.41% when involving seven empty cells and two misleading sets (R#1), in average. The seven empty cells has dragged down for 13.32% ACC in average, when they were involved in super sets swapping experiments.

- **F1-SCORE**

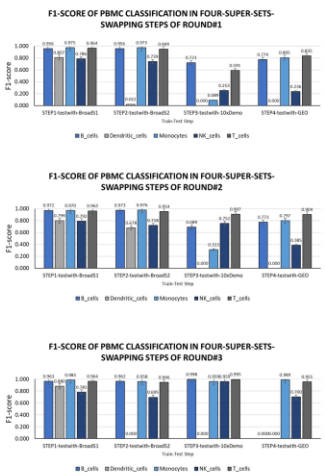


Figure 7. The F1-score results of each cell type (BC, DC, MC, NK, and TC) in each swapping train-test step of three rounds.

- After removing seven empty cells in train set, the classification ability to BroadS2 Dendritic cells have been improved for 0.656 with F1-score (R#2vsR#1).
- When excluding mislabeled data sets in train set (step3 and step4 in R#2 and R#3), F1-score of BC classification has increased from 0.689 to 0.998, MC from 0.313 to 0.959, NK from 0.752 to 0.959, TC from 0.907 to 0.995 (in terms of the large number of TC), in step3. In step4, MC has increased from 0.797 to 0.989, NK from 0.385 to 0.700, TC from 0.904 to 0.955.

CONCLUSION & DISCUSSION

- The ANN classification has been vulnerable when involving few empty cells (7/~122,600) in train set, and involving mislabeled data sets in train set.
- The correctness of label of training data sets is essential to build a solid ANN SCT classification model.
- The binary misclassification of NK and TC needs to be further studied.
- The misclassification of sub cell types (in detailed confusion matrix) needs further study with prepared SCT cell ontology.
- The comparative analysis to DC of BroadS1 and BroadS2 needs to be done. There might be underlying gene expression discrepancies between them.

PUBLICATIONS

[1] Zhong J, Lyu M, Jin H, Cao Z, Chitkushev LT, Zhang G, Keskin DB, Brusic V. Artificial Neural Networks for classification of single cell gene expression. BMC Bioinformatics. 2022.

[2] Zhong J, Shaikh RA, Wu H, Lin X, Cao Z, Chitkushev LT, Zhang G, Keskin DB and Brusic V. Classification of PBMC cell types using scRNAseq, ANN, and incremental learning. In IEEE Int. Conf. Bioinformatics Biomed. 2020, 1351-1355.

[3] Lin X, Zhong J, Lyu M, Lin S, Keskin DB, Zhang G, Brusic V and Chitkushev LT. Artificial Neural Network System for Cell Classification using Single Cell RNA Expression. In IEEE Int. Conf. Bioinformatics Biomed. 2020, 1253-1257.

[4] Shaikh RA, Zhong J, Lyu M, Lin S, Keskin DB, Zhang G, Chitkushev LT and Brusic V. Classification of Five Cell Types from PBMC Samples using Single Cell Transcriptomics and Artificial Neural Networks. In IEEE Int. Conf. Bioinformatics Biomed. 2019, 2207-2213.

ACKNOWLEDGEMENTS & CONTACT

We much thank: Guanglan Zhang and Lou T. Chitkushev, from Boston University, Derin B. Keskin, from Harvard Medical School, Dana-Farber Cancer Institute, and Zhiwei Cao, from Tongji University, for their generous help and valuable support for this research.

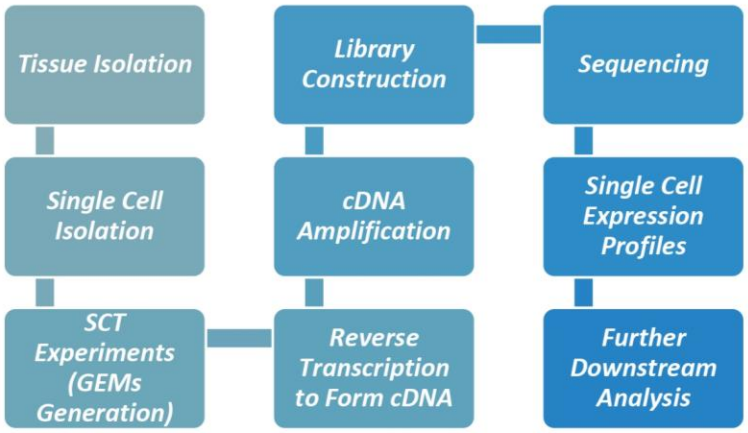
Contact information: Vladimir.Brusic@nottingham.edu.cn, Heshan.Du@nottingham.edu.cn, Huan.Jin@nottingham.edu.cn, Jiahui.Zhong@nottingham.edu.cn, 199 Taikang East Road, Yinzhou, Ningbo, China 315100.

Appendix 12 Wet Lab Background Information – Upstream Workflow and Analysis for SCT

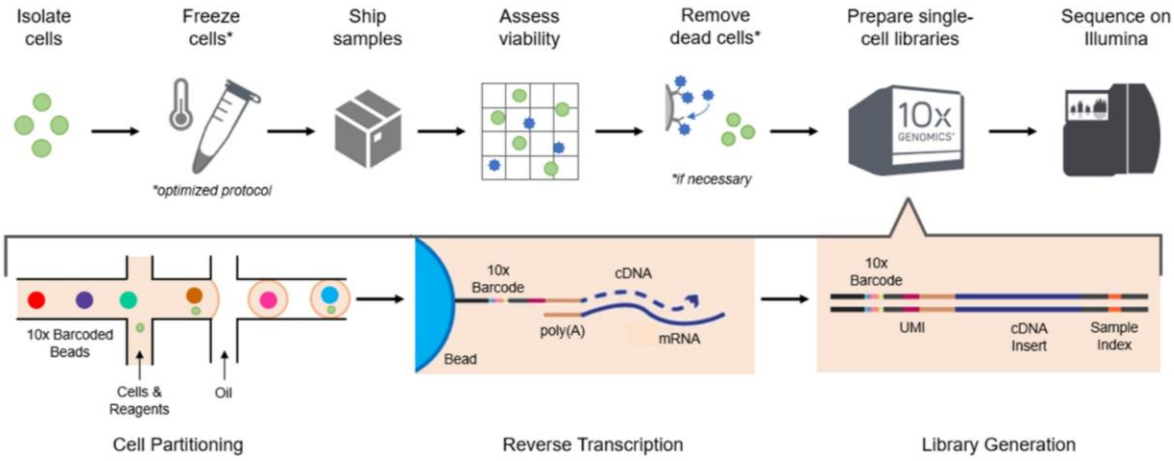
The upstream workflow and analysis for SCT can be divided into two parts:

I. *The protocol of SCT experiments (Measure transcripts and have raw data).*
II. *The upstream data analysis to the generated raw data (File conversion, alignment, QC & filtering).*

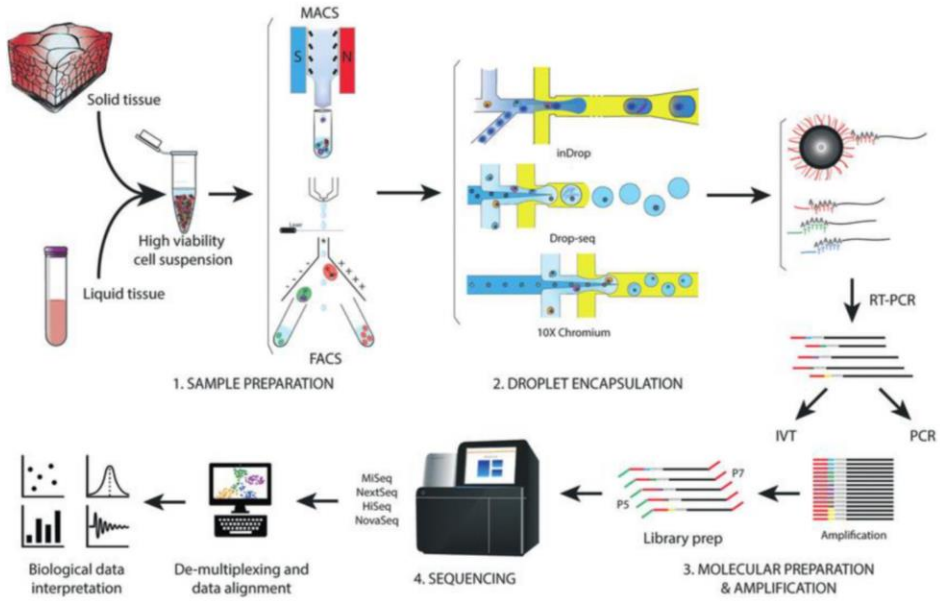
The detailed steps mainly include single-cell isolation, single-cell experiments (generation of GEMs (Gel Bead in emulsion)), reverse transcription of RNA to cDNA (emulsion PCR + barcoding), breaking GEMs, cDNA amplification, library construction and quality control, sequencing and data analysis.



The workflow of SCT sequencing technology (upstream, GEMs, 10x Genomics).



An example of SCT workflow (Azenta Life Sciences, 2023).



Another example of SCT workflow[1].

1. Single cell isolation methods

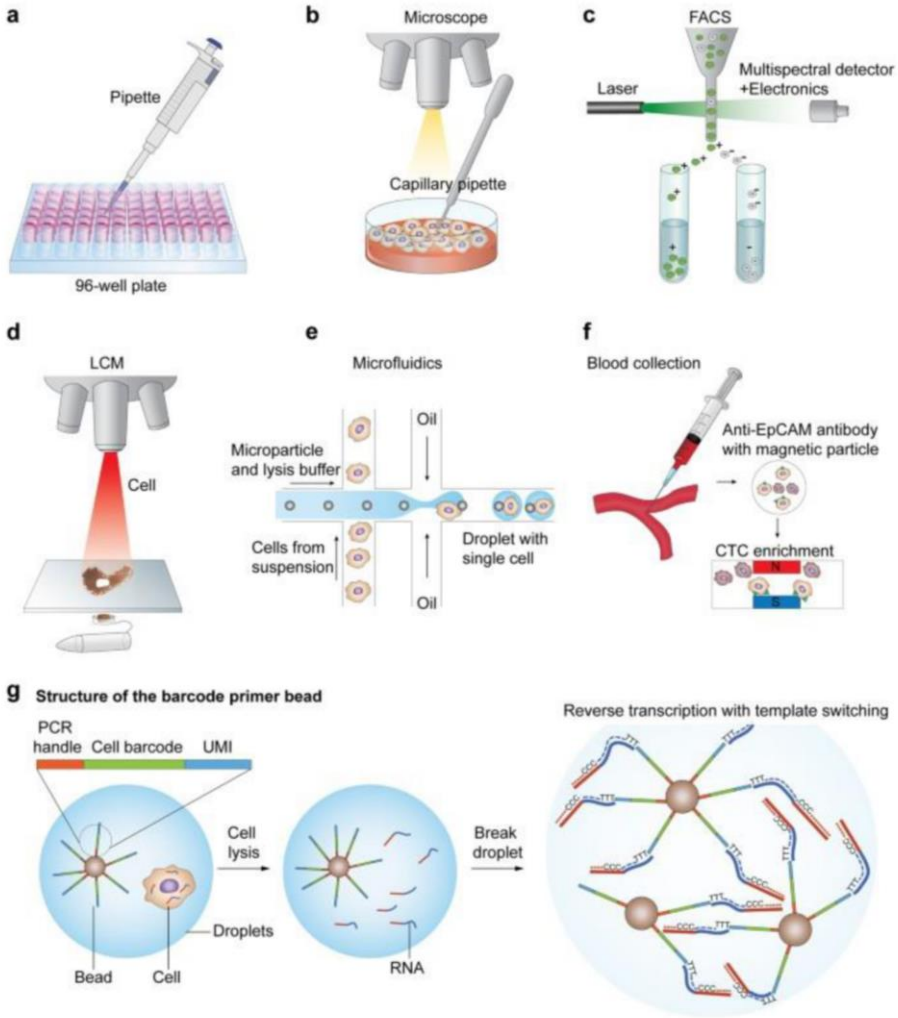
A heterogeneous population of cells must be separated into individual cells for SCT. With the availability of various current methods, when choosing a separation method, it needs to consider the experimental design requirements for cell throughput, and the requirements of the selection method - blind selection or biased selection based on a parameter.

Fluorescence-activated cell sorting (FACS) is a commonly used cell isolation/purification/sorting method, it is biased cell sorting, performed with factors such as the target surface protein markers, the same as magnetic-activated cell sorting (MACS). Unbiased cell separation is done by microfluidics (Fluidigm C1 and 10x Genomics) and droplet-based technologies (Bio-Rad ddSEQ Single-Cell Isolator). The process of tissue and cell isolation can change the profile of single-cell expression.

The manual single-cell isolation methods include laser capture microdissection (LCM) and microscope checking, they are biased selection methods based on fluorescence reporting of gene expression or cell morphology. They allow us to figure out the microtissue environment and the specific location of each single cell.

Throughput	Technologies	Cell Selection	Cell Quantity	Final Volume
Low	Microscope/LCM	Biased	$10^1 \sim 10^2$	µL
	FACS	Biased	$10^2 \sim 10^3$	µL
	MACS	Biased	$10^2 \sim 10^3$	µL
High	Microwell	Unbiased	$10^2 \sim 10^4$	nL
	Microfluidics	Unbiased	$10^2 \sim 10^4$	nL
	Droplets	Unbiased	$10^3 \sim 10^4$	nL

The comparison of different single cell isolation methods.



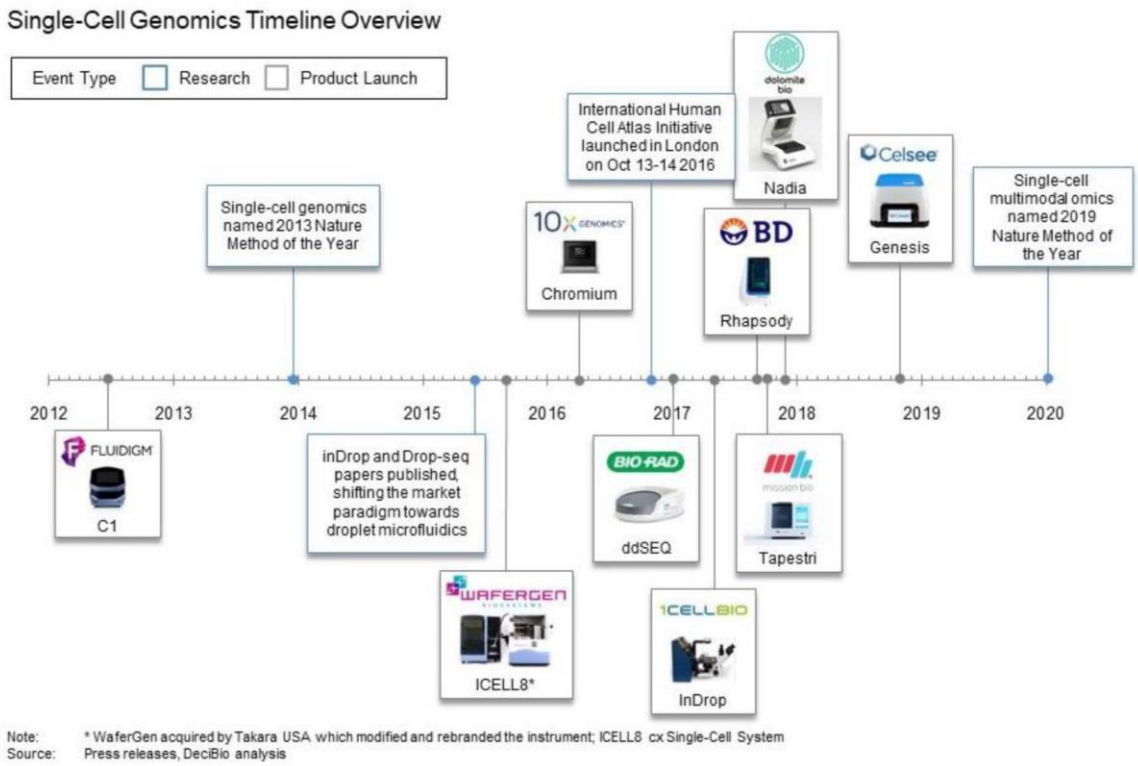
Single-cell isolation and library preparation[2].

2. SCT protocols

SCT started in 2009[3], the current mainstream SCT technologies include 10x Genomics, Fluidigm C1, and Smart-seq2.

The ideal scRNA-seq method is desired to be universal in terms of cell size, cell type, and cell state, and be cost-effective per cell, easy to use, and open source. It can assay every single cell (i.e. 100% capture rate), and detect every full-length sequence transcript in every cell (i.e. 100% sensitivity) in in-situ measurements, without doublets, minimum input of the number of cells, and additional multimodal measurements.

Currently, different SCT technologies have different advantages and disadvantages. They are selected and used according to research needs.



SCT technology timeline (DeciBio, 2021).

Comparison of single cell sequencing platforms.

SCT	Methods	Advantages	Disadvantages	Scope of Application
10x Genomics	Microfluidic-droplet	• High throughput. • Cost-efficient. • Easy to use. • High degree of automation. • Mostly used.	• High cell quantity and viability. • 3' sequencing (gene detection rate lower than full-length sequencing). • Quality control points. • High cost of personalization.	• Cells <40 µm in diameter (limited by the diameter of the instrument pipe). • Large-scale cell sample studies.
Fluidigm C1	Microfluidic capture	• Low operation requirements. • Short experimental cycle (several hours for 96 cells). • Full-length mRNA data, high gene detection rate.	• Low throughput. • High cost. • Quality control points.	• Cells 5~25 µm. • Few cell sample studies.

Smart-seq2	Manual selection	• Manual operation of cell sorting, more flexible protocol according to experimental conditions. • Low sample volume of demand. • Many quality control points, able to check cell condition from the start. • Full-length mRNA data, high gene detection rate.	• Low throughput. • High cost. • High operation requirements. • Long experimental cycle (96 cells need > one week).	• Trace cell sample studies (such as embryo cell samples, etc.).

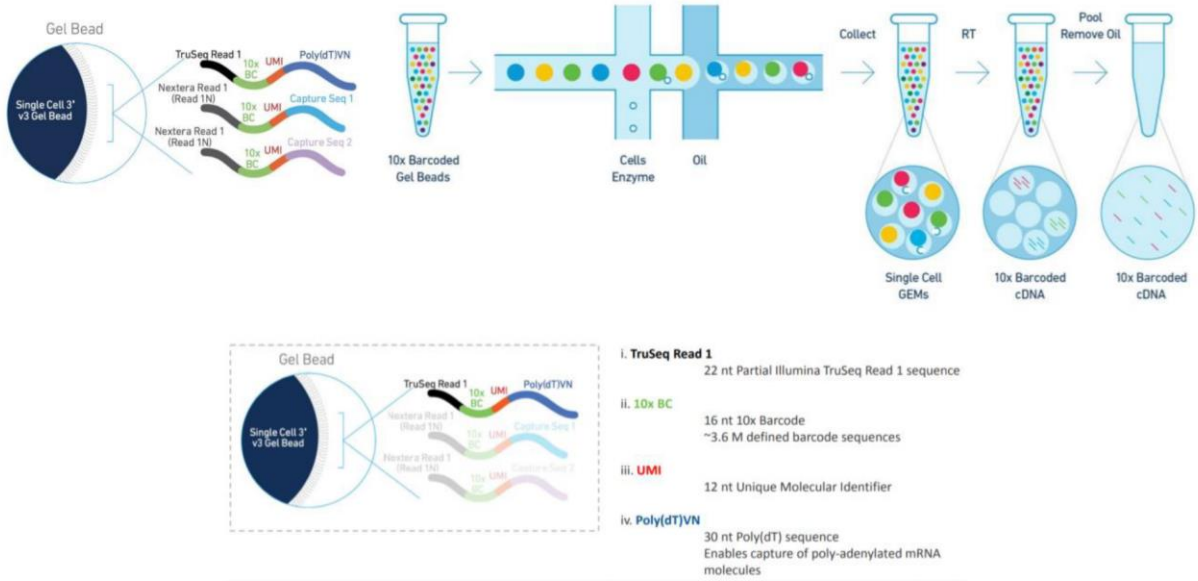
	inDrops	10x Genomics	Drop-seq	Seq-well (Honeycomb)	SMART-seq
Cell capture efficiency	~70-80%	~50-70%	~10%	~80%	~80%
Time to capture 10k cells	~30min	10min	1-2 hours	5-10min	--
Encapsulation type	Droplet 	Droplet 	Droplet 	Nanolitre well 	Plate-based 
Library prep	CEL-seq Linear amplification by IVT	SMART-seq Exponential PCR based amplification	SMART-seq Exponential PCR based amplification	SMART-seq Exponential PCR based amplification	SMART-seq Exponential PCR based amplification
Commercial	Yes	Yes	--	Yes (Summer 2020)	Yes
Cost (~$ per cell)	~0.06	~0.2	~0.06	~0.15	1
Strengths	• Good cell capture • Cost-effective • Real-time monitoring • Customizable	• Good cell capture • Fast and easy to run • Parallel sample collection • High gene / cell counts	• Cost-effective • Customizable	• Good cell capture • Cost-effective • Real-time monitoring • Customizable	• Good cell capture • Good mRNA capture • Full-length transcript • No UMI
Weaknesses	Difficult to run	Expensive	Difficult to run & low cell capture efficiency	Available Soon	Expensive

Comparison of SCT methods (HMS).

In 2017, a commercial sequencing platform (10x Genomics®) appeared, enabling single-cell sequencing technology to enter the market. The 10x platform generally provides a number of cells in the 1,000~100,000 range. This level of sequencing cell quantity can cover single-cell population types in most tissues.

In 10x Genomics, the barcoded gel beads meet and combine the cells and enzyme reagents in the first inlet of the microfluidic double-cross junction system, and then they form GEMs packaged by oil surfactants at the second inlet of the double-cross junction. Single-cell capture is achieved through this process.

10x is a reliable large-scale SCT technology and the most successful platform for commercialization so far. Currently, the vast majority of single-cell research is done with 10x technology, and the production of 10x SCT datasets has grown exponentially.



The workflow of 10x SCT technology (10x Genomics).

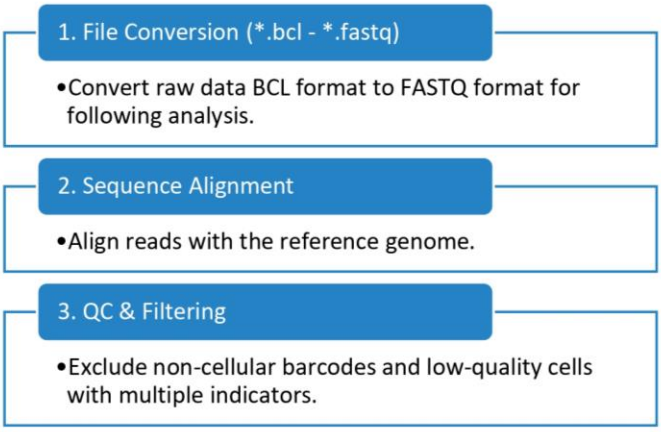
3. Reverse transcription, library construction, and sequencing

In this step, polyA selection is typically used to enrich for mRNA, and modified Oligo (dT) primers are used for reverse transcription. During reverse transcription, unique molecular identifiers (UMIs) are used to label individual molecules. Afterward, the cDNA is amplified by PCR for library construction and sequencing.

Sequencing is commonly performed on the Illumina sequencing platform. The product selection depends on the design and scale of the experiment (e.g. the NovaSeq 6000 supports large-scale studies, and the NextSeq 500 is suitable for small experiments).

4. Upstream data analysis

In general, the upstream data analysis of SCT includes three steps: 1) file conversion (base detection), 2) sequence alignment, and 3) quality control (QC) and filtering.



The general steps of SCT upstream data analysis.

1) File conversion

The raw data files produced by sequencing are in Binary Base Call (BCL) format and need to be converted to the text-based sequence file format (FASTQ) to complete subsequent data analysis.

2) Sequence alignment

It needs to map and align reads into the reference genome. It usually uses Burrows-Wheeler (BWA) aligner and STAR alignment algorithm, which aligns splice transcripts to the reference genome. The read matrix (read counts) or count matrix (gene matrix of molecular counts) (which depend on whether UMIs are used in the experimental protocol) are generated by raw sequencing data, these matrices have cell barcodes/cell numbers as the horizontal heading, gene names/gene list as the vertical heading and gene expression numbers as the digital matrix.

3) QC & Filtering

Before downstream data analysis, SCT data cell quality control needs to be done to ensure low-quality cells are removed. For example, doublets/multiplets (co-capture of multiple cells) and empty droplets (capture of no cells) can appear. This will result in the barcode incorrectly labeling multiple cells or zero cells, respectively. Read quality control (reads QC) is usually performed by assigning reads to the corresponding cellular barcode and genome expression. In the 10x protocol, this step is done with the Cell Ranger pipeline.

The QC indicators include the expressed gene number of each barcode (the number of positive), the total counts of gene expression of each barcode (the total sum of each barcode, the count depth), the ratio aligned to mitochondrial/ribosomal/hemoglobin genes, and the assessment of doublets, etc.[4]. Cells outside the standard expected range represent low-quality 'cells' that do not require downstream analysis, or they represent unusual cells that require further study. A high read ratio to mitochondria and ribosomes can be caused by increased cell

apoptosis and it can be filtered out. The number of genes that exceed the standard expectations can be used to detect and exclude doublets[5]. The QC indicators should be considered parallelly and determined coordinately, or it can lead to misunderstanding of SCT expression information[4].

The raw count matrices generally comprise 20,000~30,000 genes features. After the QC of cell states, transcript level QC also needs to be conducted by setting a threshold to filter out genes that are not expressed in most cells and won't provide valuable information about cellular heterogeneity. The setting of the threshold needs to be careful when it comes to datasets that have high dropout rates.

Cell types and states are diverse and different in datasets containing different heterogeneous cell populations, and QC strategies should be evaluated based on the results and needs of downstream analysis[4].

REFERENCES

1 Salomon, R. *et al.* Droplet-based single cell RNAseq tools: a practical guide. *Lab on a Chip* **19**, 1706-1727 (2019).

2 Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine* **50**, 1-14 (2018).

3 Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods* **6**, 377 (2009).

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

5 AlJanahi, A. A., Danielsen, M. & Dunbar, C. E. An Introduction to the analysis of single-cell RNA-sequencing data. *Molecular Therapy-Methods & Clinical Development* **10**, 189-196 (2018).

THE END