Statistical and machine learning-based approaches to precise characterization of cellular phenotypes

by

Krishan Gupta

Under the Supervision of

Dr. Debarka Sengupta

Dr. Abhik Ghosh

Dr. Gaurav Ahuja

Indraprastha Institute of Information Technology Delhi

October, 2021

Statistical and machine learning-based approaches to precise characterization of cellular phenotypes

by

Krishan Gupta

Submitted

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

to the

Indraprastha Institute of Information Technology Delhi

October, 2021

Certificate

This is to certify that the thesis titled *Statistical and machine learning-based approaches to precise characterization of cellular phenotypes* being submitted by *Krishan Gupta* to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standard fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree or diploma.

October, 2021

Dr. Debarka Sengupta (IIIT Delhi)

Dr. Abhik Ghosh (ISI Kolkata)

Dr. Gaurav Ahuja (IIIT Delhi)

Indraprastha Institute of Information Technology Delhi

New Delhi 110020

Acknowledgements

First of all, I want to express my heartfelt gratitude to my main supervisor Dr. Debarka Sengupta, for providing every necessary supervisory and emotional support to pursue my doctoral research in a focused manner within a relatively short period. He taught me about scientific pursuit, from alphabets to words and sentences. He always encouraged me to engage in scientific discussions with my peers and superiors, which led to numerous fruitful scientific collaborations in India and abroad. He inculcated numerous unheard-of best practices to ensure quality and integrity in research.

I want to thank Dr. Abhik Ghosh, my co-supervisor from Indian Statistical Institute Kolkata, for helping me with numerous mathematical formulations. I am also grateful to Dr. Gaurav Ahuja, my additional co-supervisor, for enticing me towards newer biological paradigms, which I was unaware of.

I want to thank Dr. Naveen Ramalingam, and Dr. Manisha Kumar, for providing me with valuable data to prove my computational methods experimentally. I would also like to thank Dr. Ujjwal Maulik, Dr. Sanghamitra Bandyopadhyay, and Dr. Mohit Kumar Jolly for having great collaborative work.

I want also like to sincere thank my Ph.D. monitoring committee members, Dr. Angshul Majumdar and Dr. Vikram Goyal, for their insightful comments throughout my Ph.D. journey. I would also like to thank Dr. Anirban Mukhopadhyay for his valuable feedback during my Ph.D. comprehensive exam. I would also like to thank all the reviewers for providing valuable feedback to improve the quality of my publications and the thesis.

I want also like to thank IIIT-Delhi technical and admin staff, especially Mr. Adarsh Agarwal and Mrs. Priti Patel, for being extremely helpful in the fast resolution of all technical and admin-related matters. I want to thank IIIT-Delhi faculties and staff members for providing excellent infrastructure and research environments and financial support during my Ph.D.

I heartily thank all my family members, particularly my parents, for their

continued support throughout this journey. My parents' continuous support and affection helped me to complete this journey successfully.

Many thanks to all my collaborators, colleagues, and friends, especially Manan Lalit, Shreya Sharma, Sidhant Kalra, Swagatam Chakraborti, Princey Yadav, Aayushi Mittal, Sanjay Kumar Mohanty, Vishakha Gautam, Arvind Iyer, Sidrah Maryam, Kirti Balyan, Smriti Chawla, Sarita Poonia, Priyadharshini Rai, Chitrita Goswami, Namrata Bhattacharya, Kamal Chapagai, Venkatesh Vinayakarao, Komal Rani, Aanchal Mongia, Mitali Sinha, Shiju S, Sandeep Sharma, Monalisa Jena, Aniket Das, Rahul Gangopadhyay, Anil Sharma, Sachin Yadav, Vivek Ruhela, Sumeet Patiyal, Anjali Lathwal, Neetesh Pandey, Madhu Sharma, Omkar Chandra R, Saheb Chhabra and Mukul Chhabra for providing a healthy research environment.

My Ph.D. journey was filled with sudden twists and turns, having many ups and downs. It has taught me how to deal with unforeseen situations and various responsibilities while balancing my career and personal life. Without the help of many others, the journey might not have been possible. I want to take this opportunity to thank everybody who helped in different ways to make this Ph.D. thesis possible.

Abstract

Delineation of the complex layers of biological system requires a cumulative effort from multiple disciplines of science. The present thesis work utilizes some of the interdisciplinary approaches by combining the automation and accuracy of computation to the in-depth concepts of Biology. In my thesis, I have addressed three fundamental biological problems. In one of my initial projects, I developed a computational framework by utilizing Machine Learning-based approach to build a classification model for the detection of Circulating Tumor Cells (CTCs). Moreover, I validated the authenticity of our model on a large number of publicly available scRNA-seq datasets and a newly generated CTC dataset of breast tumour cells, captured using a newly developed microfluidic system for label-free enrichment of CTCs. In my second project, I utilized single-cell genomics approach coupled with stringent statistical and structural biology frameworks to dissect the cellular basis of the loss of smell in COVID-19 infected patients. Of note, one of the prevalent, but largely ignored symptoms during the early COVID-19 pandemic was the loss of smell and taste. Our work utilized the known information about the viral entry proteins, and viral-human protein-protein interaction map. Our integrative analysis clearly suggests that the non-sensory (sustentacular, Globolar Basal Cells and Bowman's gland) cell-types are vulnerable to SARS-CoV-2 infection. In my third project, I explored the potential of modelling expression-ranks, as robust surrogates for transcript abundance. Here I examined the Discrete Generalized Beta Distribution (DGBD) performance on real data and devised a Wald-type test to compare gene expression between two phenotypically divergent groups of single cells. We carried out a comprehensive assessment of the proposed method, to understand its advantages as compared to some of the current best practice approaches. In addition to striking a reasonable balance between Type 1 and Type 2 errors, we concluded that with increasing sample size, Rank Order-Sequencing (ROSeq), the proposed differential expression test, is remarkably robust for expression noise and scales rapidly.

Publications

International Journals

1. **Gupta, K.**, Lalit, M., Biswas, A., Sanada, C.D., Greene, C., Hukari, K., Maulik, U., Bandyopadhyay, S., Ramalingam, N., Ahuja, G. and Ghosh, A., 2021. Modeling expression ranks for noise-tolerant differential expression analysis of scRNA-seq data. Genome Research, pp.gr-267070.

2. **Gupta, K.**, Mohanty, S.K., Mittal, A., Kalra, S., Kumar, S., Mishra, T., Ahuja, J., Sengupta, D. and Ahuja, G., 2020. The Cellular basis of loss of smell in 2019-nCoV-infected individuals. Briefings in bioinformatics.

3. **Gupta, K.**, Yadav, P., Maryam, S., Ahuja, G. and Sengupta, D., 2021. Quantification of Age-Related Decline in Transcriptional Homeostasis. Journal of Molecular Biology, 433(19), p.167179.

4. **Gupta, K.**, Balyan, K., Lamba, B., Puri, M., Sengupta, D., Kumar, M., 2021. Ultrasound placental image texture analysis using artificial intelligence to predict hypertension in pregnancy.The Journal of Maternal-Fetal & Neonatal Medicine (IJMF).

5. Iyer, A., **Gupta, K.**, Sharma, S., Hari, K., Lee, Y.F., Ramalingam, N., Yap, Y.S., West, J., Bhagat, A.A., Subramani, B.V. and Sabuwala, B., 2020. Integrative analysis and machine learning based characterization of single circulating tumor cells. Journal of clinical medicine, 9(4), p.1206.

6. Kalra, S., Mittal, A., **Gupta, K.**, Singhal, V., Gupta, A., Mishra, T., Naidu, S., Sengupta, D. and Ahuja, G., 2020. Analysis of single-cell transcriptomes links enrichment of olfactory receptors with cancer cell differentiation status and prognosis. Communications biology, 3(1), pp.1-10.

7. Gupta, R., Mittal, A., Agrawal, V., Gupta, S., **Gupta, K.**, Jain, R.R., Garg, P., Mohanty, S.K., Sogani, R., Chhabra, H.S. and Gautam, V., 2021. OdoriFy: A conglomerate of Artificial Intelligence-driven prediction engines for olfactory decoding. Journal of Biological Chemistry, 297(2).

8. Gautam, V., Mittal, A., Kalra, S., Mohanty, S.K., **Gupta, K.**, Rani, K., Naidu, S., Mishra, T., Sengupta, D. and Ahuja, G., 2021. EcTracker: Tracking and elucidating ectopic expression leveraging large-scale scRNA-seq studies. Briefings in Bioinformatics.

9. Gupta, A., Choudhary, M., Mohanty, S.K., Mittal, A., **Gupta, K.**, Arya, A., Kumar, S., Katyayan, N., Dixit, N.K., Kalra, S. and Goel, M., 2021. Machine-OlF-Action: A unified framework for developing and interpreting machine-learning models for chemosensory research. Bioinformatics.

International Conferences

1. **Gupta, K.**, Jain, T. and Sengupta, D., 2018, December. Texture Classification Using Deep Convolutional Neural Network with Ensemble Learning. In International Conference on Mining Intelligence and Knowledge Exploration (pp. 341-350). Springer, Cham.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Genomics is an interdisciplinary discipline of biology that specifically focuses on genome structure, function, evolution, mapping, and editing. A genome contains a complete set of genetic material (DNA or RNA) of an organism and consists of functional elements which are known as genes. A typical genome contains a coding and a non-coding region. Importantly, it has now been well established that in addition to the nucleus, almost every eukaryotic cell has organelles such as mitochondrion (plants and animals) and chloroplast (for plants) contain nucleic acid. DNA is spatially restricted within the nucleus and mitochondria in most cases, whereas, the RNA is synthesized in the nucleus or mitochondria, but could also transport to the cytoplasm. The structural backbone of DNA is the nucleotides composed of a five-carbon deoxyribose sugar backbone, a phosphate group, and a nitrogen base (Adenine, Thymine, Cytosine, and Guanine). Notably, in the case of RNA, the Thiamine base of the backbone is replaced with Uracil. An amazing fact about the human genome is that it comprises more than 3 billion DNA base pairs. Such an enormous amount of genetic material is considered the cell's functional identity, and therefore, controls all

the cellular functionalities. A cell is the smallest and functional unit of life. A cell is defined as a biological structure that contains cytoplasm and is protected by a biomembrane. Transmission of the information from the cell nucleus to the cytoplasm is explained using the Central Dogma of Molecular Biology. In the central dogma, DNA replicates, and also information in genes translates into proteins. It consists of two phases: (1) transcription and (2) translation. In the transcription phase, a segment of DNA encodes for RNA with the help of RNA polymerase. In the translation process, the messenger RNA (mRNA) molecule sequence translates to a sequence of amino acids. This process is called translation or protein synthesis. The genetic code in DNA describes the relationship between the sequence of nitrogen bases within a specific gene and the corresponding amino acid sequence that encodes the associated protein **(Figure 1.1)**.

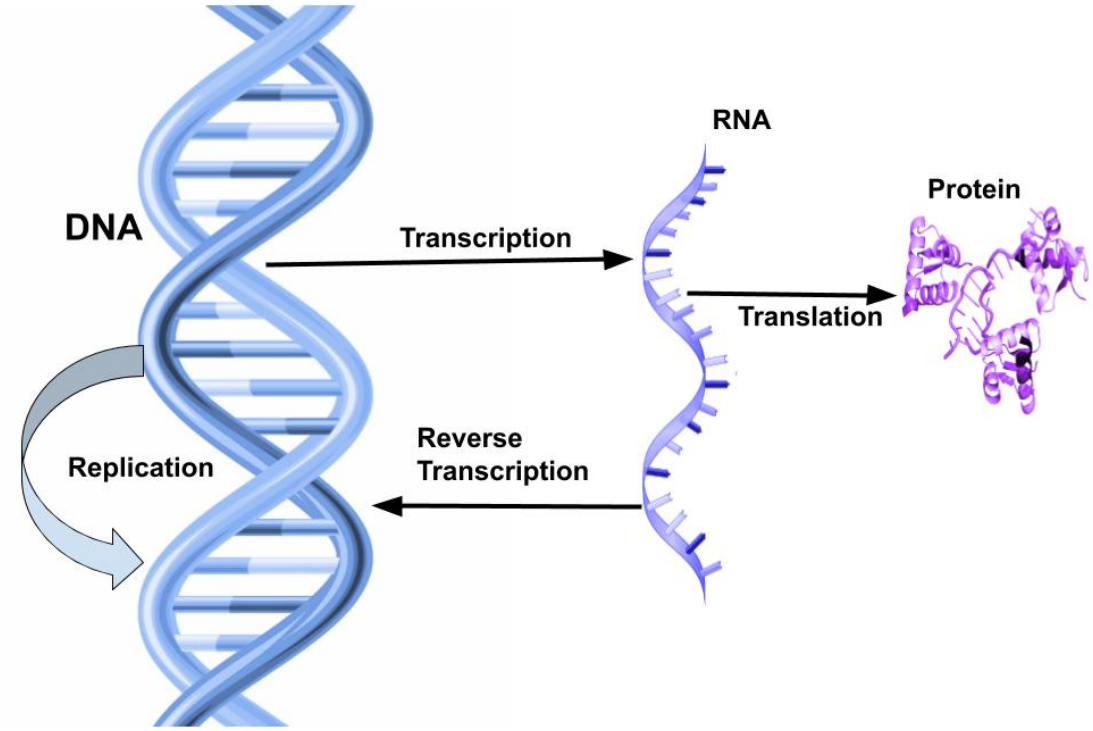

Figure 1.1: The central dogma of molecular biology and its various processes, including transcription, translation, reverse transcription, and replication.

In eukaryotic cells, transcription occurs within the nucleus, and then the mRNA transports from the nucleus to the cytoplasm through the specialized pores within the nuclear envelope. Once the mRNA reaches the cytoplasm, it allows the translation of these processed RNAs into proteins/polypeptides via ribosomes, a specialized cellular organelle involved in translation. In the case of a prokaryotic cell, however, it lacks the proper nuclear membrane, and therefore transcription and translation processes occur simultaneously. Translation begins while the mRNA is still transcribing, leading to the concept of co-transcription/translation. The sequence level information about any DNA or RNA molecule can be obtained using the Sanger sequencing. It is a well-established, Nobel prize-winning technique used to get sequence-level information. The technique was established in 1977 by two-time Nobel Laureate Frederick Sanger. In comparison to the Sanger sequencing, the Next-Generation Sequencing (NGS) allows researchers to perform large-scale whole-genome sequencing (WGS) with minimal hassle. Analyzing or extracting valuable information from these gigantic datasets requires computational tools and methods. Computational biology is an interdisciplinary field that allows the development of algorithms, methods, and models to understand the complex biological systems and the relationships between them. Usage of Machine and Deep Learning in Biology eases the entire process, as they provide the much-needed and most simplified solution to decode this complex data. The major areas which require immediate attention are developing robust methods for the Differentially Expressed Genes, improving Clustering and Classification of the cells/genes/RNA. Notably, with the advancements in the omics techniques (genomics, transcrip-

tomics, proteomics, and metabolomics) the overall requirement of the computational methods and tools are ever-increasing and therefore provides an opportunity to obtain the systematic molecular view of the underlying biological processes.

1.1 Background

Phenotype and genotype, these two readouts have traditionally been used by researchers to assess the relationship between a genes and its functionalities. For all cellular processes, DNA is the underlying blueprint, RNA is the molecule produced on demand when certain processes are required. DNA is invaluable for studying heritable disorders and allowing statistical evaluations of a biological sample, but it tells so little about a cell's complex, real-time behaviours. Researchers have been using RNA sequencing (RNA-seq) to relate gene expression and physiological conditions. RNA-seq identifies the genes in a cell that are turned on, and the intensity at which they are expressed. This helps scientists to grasp a cell's genetics thoroughly to determine modifications that can signify illness. In recent years, RNA-seq has emerged as a powerful technology for transcriptome profiling. A transcriptome is a global overview of the cellular RNAs.

1.1.1 RNA species

RNA is a polymer of nucleotides, made up of phosphate, ribose sugar, and bases (adenine, cytosine, guanine, and uracil). In both eukaryotic and prokaryotic

cells, there exist three main types of RNA. First is messenger RNA (mRNA) as the most heterogeneous form of RNA that accounts for just 5% of the total RNA of the cell. It carries complementary genetic code copied from DNA in the form of triplets of nucleotides called codons. Second, ribosomal RNA (rRNA) accounts for 80% of the total RNA in the cell. Different rRNAs are found in the ribosomes, including small and large rRNAs. Third, transfer RNA (tRNA) is found with the cloverleaf structure due to strong hydrogen bonding between the nucleotides. Some other types of RNAs are small nuclear RNA (snRNA), transfer-messenger RNA (tmRNA), ribozymes (RNA enzymes), double-stranded RNA (dsRNA), regulatory RNAs, including antisense RNA (aRNA), small interfering RNA (siRNA), and micro RNA (miRNA). In the Central Dogma of Molecular Biology, the mRNA lies at the intermediary stage between the genetic material, i.e. DNA and the proteins. mRNA contains codons and directs the formation of amino acids through ribosomes and Transfer RNA (tRNA). mRNA contains multiple regulatory regions that can determine the timing and rate of translation. It also ensures translation proceeds in an orderly manner as it has sites for the tRNA, docking of ribosomes, as well as various helper proteins.

1.1.2 Estimation of mRNA expression

RNA sequencing (RNA-seq) is a widely used genomics approach for estimating and analyzing quantitative messenger RNA molecules in a biological sample, thereby studying the cellular responses. RNA-seq has fuelled much innovation and discovery in medicine over recent years. Estimation of mRNA expression

is important because direct measurement of protein expression levels is still challenging. Various methods to estimate mRNA or cDNA expression levels have been developed, including microarray [5, 6, 7], polymerase chain reaction (PCR) amplification [8], sequential analysis of gene expression (SAGE) [9], Expressed Sequence Tags (EST), abundance [10, 11, 12, 13] etc. Gene expression can also be analyzed by directly measuring protein levels with a technique known as Western Blot [14]. Two other most popular techniques are Northern Blot, and Serial Analysis of Gene Expression (SAGE), which also pinpoint the actively transcribing genes within a cell [15, 16].

1.1.2.1 Bulk expression

In complex tissues consisting of multiple distinct cell types, bulk RNA-seq measures the average gene expression levels by summing over the cellular population. As a result, variability in cell-type compositions confounds analysis such as detecting differential gene expression [17]. Before the advent of single-cell RNA sequencing, bulk RNA sequencing was the method of choice for profiling transcriptomes under various conditions such as healthy, demographic, disease, infectious, etc. [18, 19, 20]. Various statistical and computational methods have been developed for bulk RNA-seq data. Still, most of these have some limitations, like they demand a priori knowledge, information about the cell-type compositions, either of gene expression profiles of purified cell types or pre-selected marker genes [21, 22, 23, 24]. In recent times, Next-generation Sequencing (NGS) based approaches are used to estimate the RNA quantities in biological samples at a given time. It allows analyzing the continuously changing

cellular transcriptome [25]. It also facilitates the ability to look at spliced transcripts, mutations/SNPs, post-transcriptional modifications, and gene expression changes at different times, groups, or treatments. RNA-Seq can also look at diverse RNA populations, including total RNA, miRNA, tRNA, and ribosome profiling [26, 27]). Prior to the advent of RNA-Seq technologies, gene expression studies were done with hybridization-based microarrays. There were various issues associated with this approach such as poor quantification of highly and lowly expressed genes, cross-hybridization artifacts, and the need to know the sequence a priori [28]. Due to these technical issues, transcriptomics transitioned to sequencing-based methods, first Sanger sequencing of expressed sequence tag libraries, then chemical tag-based methods, and finally to the current technology, RNA-Seq.

1.1.2.2 Single-cell expression

The single-cell analysis allows the scientists to study cell-to-cell variation within a cell population like an organ, tissue, cell types, etc. Single-cell sequencing is the NGS technology that examines the sequence information from individual cells, thus providing a higher resolution of cellular differences and a better understanding of an individual cell's functionality in the context of its microenvironment. Single-cell transcriptomics enables exploring the gene expression level of individual cells simultaneously by measuring their RNA content. In recent times, 10X genomics provide software suite named 'Cell Ranger' to process Chromium single-cell RNA-seq output by allowing reads alignment and generation of feature-barcode matrices. The Cell Ranger workflow can begin

with demultiplexing the raw base call (BCL) files for each flowcell directory, FASTQ files that have already been demultiplexed with 'bcl2fastq' directory etc.

1.1.2.3 Spatial transcriptomics

Spatial transcriptomics is a groundbreaking technology that gives the relationship between cells and their locations. The cellular site allows scientists to measure all the gene activity in a piece of tissue and map where the event is occurring and can be critical to understanding disease pathology. Spatial transcriptomics technology also provides the relative location of RNA concerning the centroid of the corresponding cell that also allows scientists to find heterogeneity within cells of the same cell types alongside exploring the area of research in the study of RNA density in a particular region within the cell. Spatial transcriptomics is used to spatially resolve RNA-seq data in individual tissue sections [29]. The barcoded primers bind and capture adjacent mRNAs from the tissue during the attachment of tissue cryosection to a spatial transcriptomic slide. At the time of attachment of tissue section to the slide, reverse transcription of captured mRNA is initiated, and then the resulting cDNA incorporates the spatial barcode of the primer. Further sequencing libraries are prepared and analyzed with Illumina dye sequencing after mRNA capture and reverse transcription. The spatial barcode of each generated sequence allows individual mRNA transcripts to be mapped back to their point of origin within the same tissue section **(Figure 1.2)**.

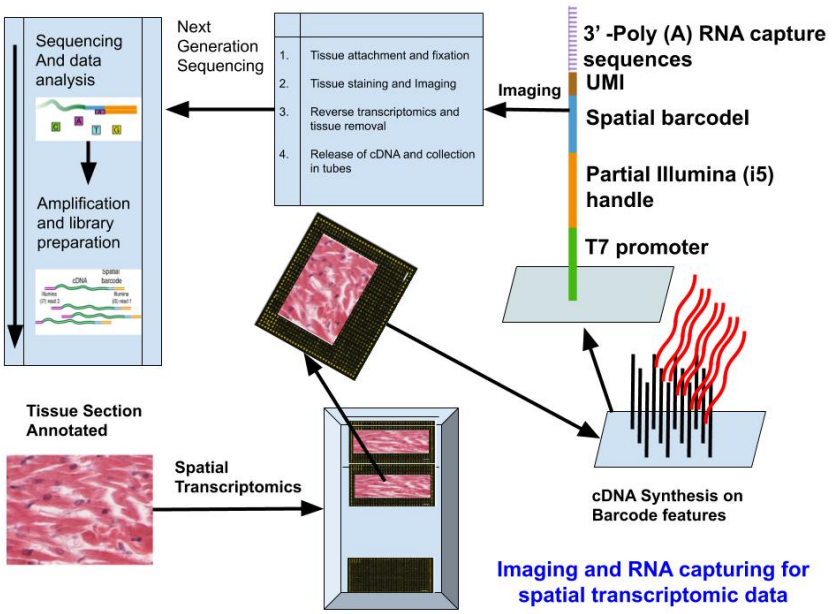

Figure 1.2: Spatial transcriptomics pipeline, capturing image and RNA estimates within cells of a particular section of tissue.

1.1.3 Noise and bias in single-cell expression estimates

In recent times, single-cell expression profiling has become mainstream in explorative studies in diverse fields of biology. Although it provides the best possible resolution into a biological system, single-cell protocols suffer from noise and technical biases due to inadequate starting materials and cell to cell variability. [1, 30]. Current scRNA-seq technology requires amplification of the amount of mRNA present in an individual cell within a minute to prepare next-generation sequencing libraries. That contributes to a substantial increase in technical noise level relative to bulk-level RNA-seq [31, 32, 33]. Despite improvements in measuring technologies, various technical factors like cell cycle effects, library size differences, and low RNA capture rate leads to massive noise in scRNA-seq experiments [34, 35, 36]. Recent droplet-based scRNA-seq technologies are sparse due to relatively shallow sequencing [37]. These techni-

cal factors produce a lot of noise, which may corrupt the underlying biological signal and obstruct analysis [38].

1.1.4 Application of machine learning in single-cell genomics data

Machine learning is an emerging field in artificial intelligence that provides systems with the ability to learn and develop automatically from experience without explicit programming. It explores the study of computer algorithms that improve automatically through experience. Three broad categories of machine learning are supervised, unsupervised, and reinforcement learning. Machine learning approaches also include ranking, active learning, representation learning, and transfer learning. Many methods have been developed from time to time to perform popular tasks like classification, clustering, dimension reduction, noise filtering, etc. Machine learning approaches have recently demonstrated their potential to automatically process and learn from large amounts of high-dimensional data in fields such as computer vision and natural language processing. The increasing adoption of high-throughput single-cell omics technologies in biological studies has created an urgent need for efficient and commensurate computational, statistical and machine learning-based methods. Some broad application areas for such methods include:

1. Detection and identification of biomarkers in the cell sub-populations of specific interest.

2. Integration and batch correction of single-cell datasets generated in multiple wet-lab environments.

3. Dimension reduction of single-cell data by discarding unimportant genes, and technical artifacts.

4. Clustering of single cells to discover heterogeneity within seemingly similar cells.

5. Classification and annotation of groups of cells to infer cell lineages.

1.2 Biological insights through single-cell expression studies

Appreciably more technologies have been developed since the first scRNA-seq study, which was published in 2009. The rapid ongoing maturation of bioinformatics approaches and the commercial availability of scRNA-seq platforms have enabled biologists to explore healthy and disease tissues at unprecedented resolution. Previous scientific studies in cell biology were mostly limited to data generated by bulk profiling methods as the only estimate averaged read-counts that generally mask cellular heterogeneity. However, this averaged approach is problematic in the study of only a subpopulation of cells such as stem or progenitor cells within a particular tissue or immune cell subsets infiltrating a tumor. Transcriptomics has recently emerged in the development of single-cell as a powerful tool to investigate cellular heterogeneity at individual cells' resolution.

1.2.1 Tissue heterogeneity

The tissue is a group of cells that perform a specific task together. The human body has mainly four types of tissues: epithelial, muscle, connective, and

nervous tissue. Blood is a form of connective tissue made up of blood cells and platelets. Tissue heterogeneity is defined as unintended profiling of cells of other origins compared to the profiling of target tissue, a common source of variance that inflames irreproducibility. Tissue heterogeneity means different rates and patterns of growth in adjacent tissue regions, unlike homogeneous growth where a region expresses a uniform rate or pattern of growth.

Single-cell studies have valuable tools for dissecting cellular heterogeneity in complex systems [39]. Human cell landscape (HCL) determines the cell type composition of major human organs and uncovers a single-cell hierarchy for many tissues that have not been well-characterized [40]. In HCL, more than 700,000 single cells are analyzed from more than 50 human tissues and cultures; generally, 2-4 replicates per tissue. HCL provides a complete human tissue dataset of 102 significant clusters.

1.2.2 Developmental trajectories

Trajectory inference (TI) or pseudotemporal ordering is a computational technique to determine the pattern of a dynamic process experienced by single cells and then arrange cells based on their progression through the process. Single-cell RNA-seq has revolutionized modern biology by allowing scientists to profile transcript abundance at the resolution of an individual cell. It explores new research insights to study cellular pathways during the cellular activation, cell cycle, and cell-type differentiation. scRNA-seq can provide a snapshot of the transcriptome of thousands of single cells in a cell population, with each cell at distinct points of the dynamic process. This advancement of transcriptional

information is, however, associated with many data analysis challenges. Statistical and computational efforts have focused chiefly on trajectory inference (TI) methods, a wide range of which has been proposed, 45 of which are extensively bench-marked [41]. Here the term trajectory refers to the collection of lineages for the process under study; in little detail, first, allocate cells to lineages and then order them based on pseudoprimes within these lineages. Profiling of Gene expression has been widely applied to understand the regulation of cellular processes in development and disease management [39].

1.2.3 Cancer heterogeneity

The cells in a body all have specific tasks to perform. A cell usually divides in an orderly and controlled manner. Cells also have individual life, they die when damaged and new ones are manufactured. Tumor cells are formed whenever a cell or a group of cells grow abnormally. This kind of abnormal growth can be time-specific or uncertain. This abnormal growth stops with the formation of tumor cells in the time-specific case; otherwise, it comes up with cancerous cells. The cancer cells undergo uncontrolled proliferation. They can grow in any part of the body and induce tissue dysfunction at the new site. Cancers are alike in many ways but are differentiated based on their growth speed and ability to spread other organs. Cancer cells can circulate in the body with blood cells through blood vessels and affect various organs, including the host. The circulation of tumor cells in the body is known as metastasis. The tumor cells circulating in the entire body through blood vessels after detaching from the primary site are circulating tumor cells (CTCs) **(Figure 1.3)**.

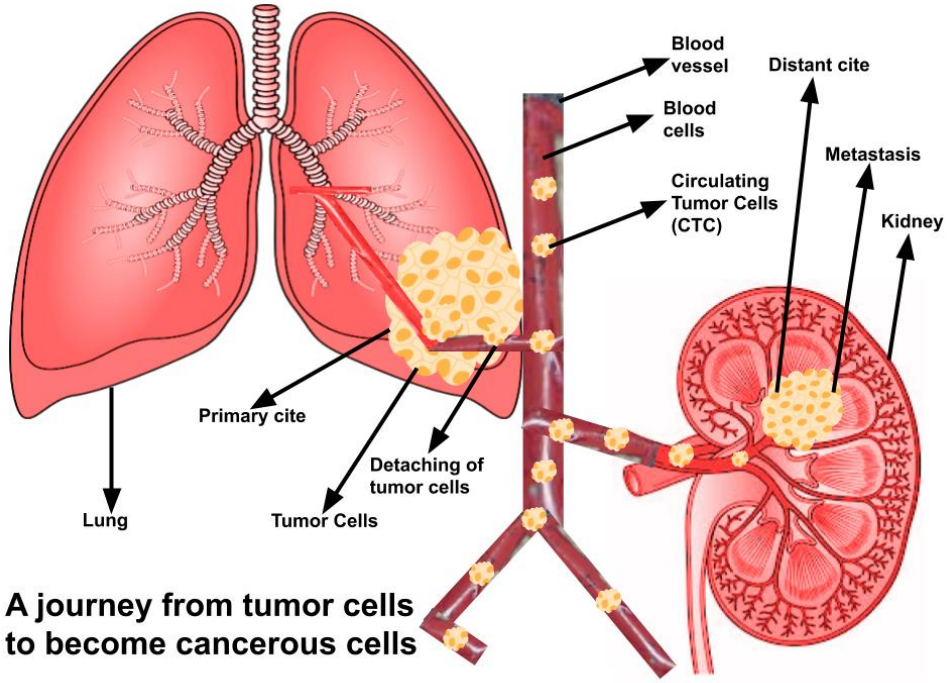

Figure 1.3: A journey of tumor cells to become cancerous cells including metastasis.

The recent advent of scRNA-seq has allowed molecular profiling of single CTCs [42], captured using microfluidic devices [43, 44, 45, 46, 47]. CTCs provide unabated, real-time insights into tumor development and therapeutic responses as a blood-based biomarker. The rareness of CTCs in the peripheral blood hinders their isolation and characterization despite these commitments [48]. Tumor heterogeneity is the existence of subpopulations of cells with distinct phenotypes and genotypes that may harbor divergent biological behaviours. Tumor heterogeneity explores the distinct morphological, and phenotypic profiles in the different tumors, including gene expression, proliferation, cellular morphology, metabolism, motility [49]. Both types of heterogeneity (intra-tumor and inter-tumor) are responsible for this phenomenon. Minimal intra-tumor heterogeneity refers to the simple consequence of the imperfection of DNA replication. One or more mutations are acquired when the cells (healthy

or cancerous) divide leading to a diverse population of cancer cells [50]. Heterogeneity of cancerous cells introduces significant challenges related to designing effective treatment strategies. Advanced research and innovation into understanding and characterizing heterogeneity can be useful in the identification of the causes and progression of the disease [51]. Heterogeneity in cancerous cells is not limited to the differences between patients and can also occur within a single patient.

1.2.3.1 Single-cell studies in tumor heterogeneity and metastasis

Heterogeneity is predominant in human cancer and presents itself as morphological variations within cells, or distinct karyotypic patterns, rates of expression of proteins and biomarkers, and genetic profiles [52, 53]. Temporal and spatial heterogeneity are results of tumor progression. Multiple tumor copies characterize spatial heterogeneity in different regions of the primary tumor or various tissues of the same patient with metastatic lesions [54]. Temporal heterogeneity results from localized to metastatic disease progression are characterized by an intra-patient heterogeneity between the primary tumor and metastatic lesions. For example, patients with $HER2_{-}$, also known as $ERBB2_{-}$ primary breast carcinomas, can present with $HER2_{+}$ metastases [55].

Metastasis is the cause of most patient deaths, and it is challenging to clinical diagnosis and experimental research. The general dogma is that rare cells undergo metastasis with distinct molecular and cellular properties [56]. Tumor cell detection in peripheral blood explores a unique non-invasive opportunity to understand and investigate how a particular step in the tumor dissemination pro-

cess can contribute to tumor heterogeneity. CD44 is a stemness marker that mediates intercellular interactions of $CD44-CD44$ haemophilic protein results in cell cluster formation. Subsequent $CD44-PAK2$ interactions activated FAK signalling; another molecular pathway reported playing a role in cancer stem cells [57].

Phenotypic and genotypic characterization of CTCs at the single-cell level shows a substantial heterogeneity that better understands the biology of tumor evolution. Emerging single-cell technologies explore individual cells' profiles within tumors and investigate the distinct genetic and phenotypic properties that can differentially promote progression, metastasis, and drug resistance. Single-cell investigations now enable these cells, including their position in primary tumors, to be recognized and characterized. The impact of genetic versus non-genetic and intrinsic versus extrinsic factors on metastasis [58] **(Figure 1.4)**.

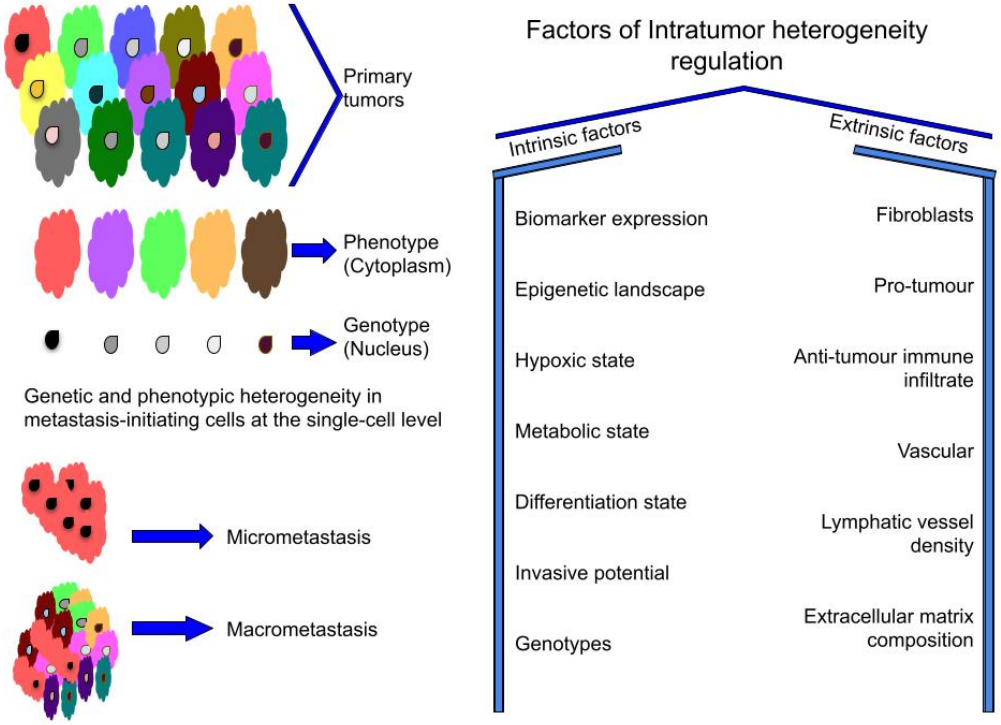

Figure 1.4: Tumor heterogeneity and metastasis with distinct molecular and cellular properties.

1.2.4 Characterization of CTCs

A circulating tumor cell (CTC) is a cell that has shed into the vasculature or lymphatic's from a primary tumor, detached from primary cites, and is carried around the body in the blood circulation [59]. CTCs might extravasate and become seeds for the subsequent growth of new tumors through metastases in distant organs responsible for the vast majority of cancer-related deaths. CTC can be circulated in cluster form; these kinds of CTC are powerful to metastasis. There is one FDA-approved method for CTC detection in recent time, Cell Search, which is used to diagnose prostate, breast, and colorectal cancer [60]. The molecular characterization of CTCs is fundamental to the description of the relevant genetic alterations and phenotypic identification of malignant cells. Molecular characterization may change according to disease progression and therapy resistance. Molecular characterization of CTCs is very challenging because of their rarity (hardly 1 CTCs in millions of cells), heterogeneity of CTCs, and technological difficulties in the enrichment, isolation, and molecular characterization of CTCs. Array technology (DEPArray) identifies each cell population entity based on multiparametric fluorescence and bright field criteria. Due to its capacity, qRTPCR on single cells can be used to define the phenotype of each isolated CTC [61, 62]. CTCs exhibit cytoplasmic expression of cytokeratin express EpCAM and CD45-negative and contain a nucleus that binds to the nucleic acid dye 4', 6-doamidino-2-phenylindole. The absence of one of these characteristics disqualifies a cell as a CTC [63] **(Figure 1.5)**.

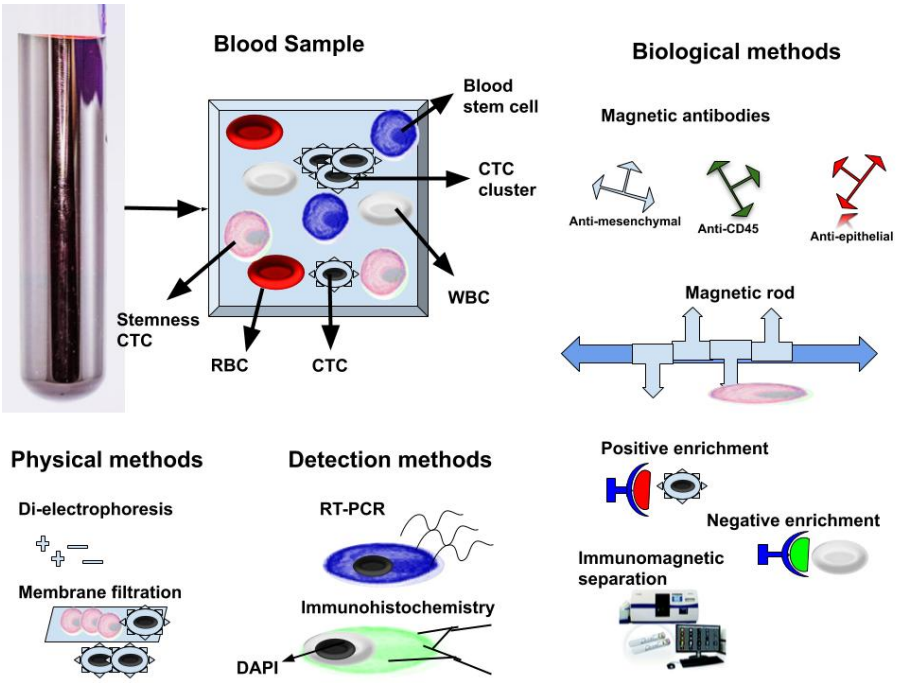

Figure 1.5: Various methods to isolate CTC from blood cells.

1.2.5 Mono-allelic expression

Monoallelic gene expression (MAE) describes the phenomenon of the gene expression where only one of the two alleles is transcribed (actively expressed), while the other is silent [64]. Alleles, in this context, mean two or more gene copies. Diploid organisms carry two homologous copies of each chromosome, and a gene can be expressed from monoallelic expression (single chromosomes) or biallelic expression (both chromosomes). MAE is of two types. The first is a random monoallelic expression (RME) and the second is the constitutive monoallelic expression (CME). RME is a broader class of monoallelic expression, defined by random allelic choice in somatic cells, to allow different multicellular organisms to express different alleles. CME occurs from the same specific allele throughout the whole organism or tissue due to genomic imprinting [65].

1.3 Some of the key computational challenges

The machine learning approach is recommended to have at least more samples than the number of features. One major challenge in gene expression statistical analyses is the small number of replicates accompanied by a large number of gene properties in data. Various kind of noise occurs during single-cell RNA-seq profiling and library preparation for same. Some of the noises are technical, cellular, due to high dropout, library preparation differences, amplification, etc. The first key challenge is to discover and innovate robust methods to analyze single-cell data. The second challenge is batch correction while the same kind of data coming from different environments. Many data sets are publicly available, but accessing data from multiple sources requires integration and batch correction. Data with these all types with noises are more challenging to come up at the same platform, as no robust method is available in recent times.

1.3.1 Challenges in differential expression analysis

One primary reason that analyses scRNA-seq data challenging is dropouts. In dropouts, the data only captures a small fraction of each cell's transcriptome; in the past decade, high-throughput microarrays have been the predominant technology for measuring gene expression. Individual array measures the expression levels of all genes from one sample. Despite its multiple arrays, the variation of the gene expression levels is captured from different samples. Some of the variations are due to technology, and some are biological variation [66]. Genetically identical cells with the same tissue are often observed to have dif-

ferent expression levels of various sizes, structures, and proteins. The reason for random variability in quantities arising in cellular biology is known as Cellular noise. [67, 68]. Cellular noise is of two types, intrinsic and extrinsic noise. Extrinsic noise is variation in identically-regulated quantities between different cells or cell-to-cell variation in a given gene expression. Intrinsic noise is variation in identically-regulated quantities within a single-cell or intra-cell variation in expression levels of two identically controlled genes. In scRNA-seq, technical parameters that describe the amplification bias and the dropout rates should be cell-specific to adjust the possible presence of systematic differences across cells. Gene expression can be significantly increased in captured cells from the sites with large or small plate output IDs in data generated by the Fluidigm C1 platform [69].

1.3.2 Cellular heterogeneity analysis and clustering

A homogeneous cell population can show cellular heterogeneity due to the influence of different internal and external stimuli [70]. Classification of these distinct cell types from the cell population has a significant contribution to biological research. There is no such standard computational method for cell-type identification. Gene expression analysis of single cells has explored the potential of revealing lots of new information about cell type and functionality. Since many past clustering techniques like Dropclust, Seurat, Zheng, et al., etc., have been developed to analyze the Gene expression of single cells [71, 72, 3].

1.3.3 Pseudotemporal analysis and RNA velocity: Decoding development

In development biology, Multi-cellular organism development begins from a single-celled zygote that undergoes rapid cell division to form the blastula. The quick, multiple cell division rounds are called cleavage. When cleavage produced more than 100 cells, the embryo is called the blastula. Human fertilization is the fusion of a human egg and sperm, usually in the ampulla of the Fallopian tube. The result of this fusion is the production of a zygote cell. The fusion of only one sperm with one egg ensures the correct number of chromosomes in the zygote. In the blastula phase, cells are known as un-committed cells. Un-committed cells can go in distinct cell types, which significant challenge to discover the cell types of an un-committed cell. In the blastula, cells rearrange themselves spatially to form three layers of cells; each layer is called a germ layer, which differentiates into different organ systems. This whole process is known as gastrulation, where the blastula folds upon itself to form the three layers of cells. During development, differentiation happens on a timescale of hours may be of days, which is typically equivalent to the half-life of mRNA. The relative abundance of spliced and unspliced mRNA can be exploited to estimate gene splicing and degradation rates without the need for metabolic labelling [73, 74]. Single-cell RNA sequencing can release RNA abundance with high quantitative accuracy, sensitivity, and throughput. RNA abundance is a reliable indicator of the state of individual cells [75]. In simple words, RNA velocity analysis of single-cell data using spliced and unspliced transcripts to better understand how a cell migrates from one state to another and in what time frame.

1.3.4 Cell-types vulnerable to infection - how it is important in COVID-19

Cells are the primary and smallest functional units of leaving objects. Cells can survive by themselves, for example, bacteria or archaea, or as part of a multi-cell organism, for example, cells of animals. Viruses come under non-living microscopic infectious parasites, generally much smaller than cell-like bacteria, and cannot thrive and replicate outside of a host body, always required a living host to replicates. Viruses have a reputation for being the cause of contagion, and widespread disease and death events have bolstered such a status. The Russian scientist Dmitry I. Ivanovsky and the Dutch scientist Martinus W. Beijerinck studied the biological nature of viruses in 1892 and 1898. Beijerinck first identified that the virus is an alive, reproducing organism that differed from other organisms; under a study, it was a new infectious agent. Their study found that an agent could transmit a disease of tobacco plants; passing through a minute filter would not allow the passage of bacteria, later called tobacco mosaic virus. In the 2014 Ebola outbreak, in 2009 H1N1/swine flu pandemic, and recently in 2019 outbreak of the COVID-19 pandemic (a widespread global outbreak) likely come to mind. Viruses can infect any leaving objects, from animals to microorganisms, including bacteria, even plants. In recent times, a sub microscopic infectious agent named as 2019-nCov spreads in whole worlds, infecting almost all countries.

Spreading of the virus occurs due to its reproductive cycle inside and outside of the body. The cycle includes infection of a cell, generates multi-copies of itself, then infection of other cells, in more detail with an example of a re-

cent virus named '2019-nCOV' that is cause for a widespread global outbreak. SARS-CoV-2 can enter inside the body via many sources like mouth, nose, eye, etc. Further, it can bind with the cell receptors like Angiotensin-converting enzyme 2 (*ACE2*) that allows its entry inside the infected cells. Further inside the cell, SARS-CoV-2 creates multi-copies of itself. After creating sufficient copies, new viruses (copies) come out from the infected cells by destroying their plasma membrane. Other new viruses can affect other cells in the same way or might also come out from the body to affect other living things **(Figure 1.6)**.

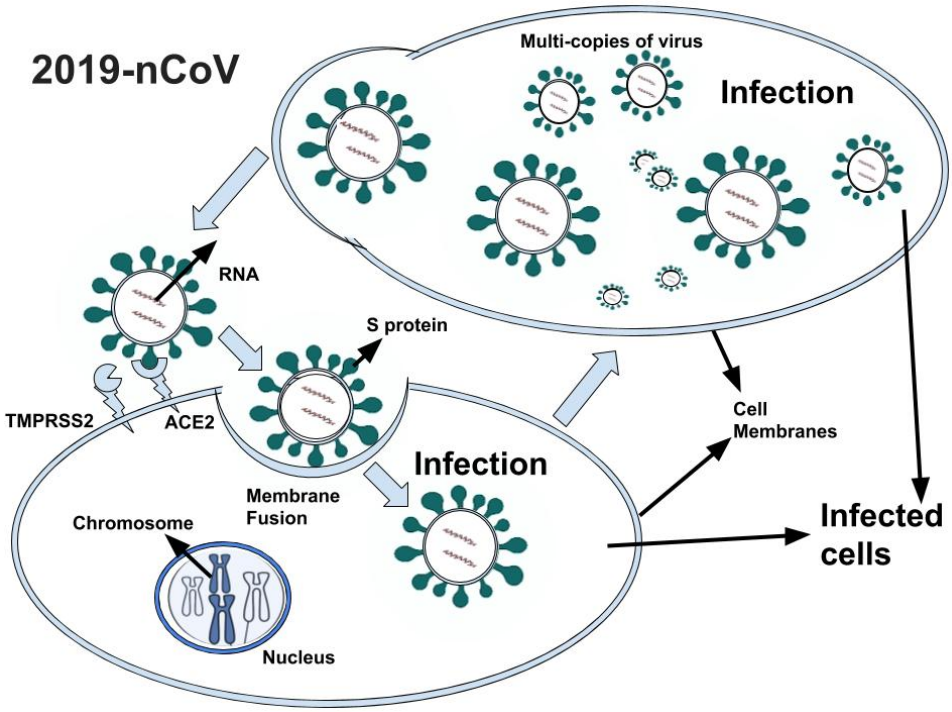

Figure 1.6: 2019-nCoV infection and its cycle to generate multi-copies in multi-cells of body.

ACE2 is an enzyme attached to the membrane of cells in the lungs, heart, kidney, etc., also obey as the entry point into cells for some coronaviruses, including HCoV-NL63, SARS-CoV, and SARS-CoV-2 [76]. *TMPRSS2* activates some of the viruses including the SARS coronavirus of 2003 and the SARS-CoV-2 and can thus be inhibited by *TMPRSS2* inhibitors [77]. SARS-CoV uses

the receptor *ACE2* for entry and the *TMPRSS2* for S protein priming [78]. So, the major problem comes with how to prevent its spreading from one leaving thing to another. Researchers worldwide are trying to understand exactly how the virus spreads, avoid transmission and develop a vaccine. It is known that the virus SARS-CoV-2 uses a similar mechanism to infect our cells as a related coronavirus that caused the 2003 SARS epidemic. But the exact cell types involved in the nose had not previously been pinpointed. Receptor protein *ACE2* and the *TMPRSS2* protease that can activate SARS-CoV-2 entry are expressed cells on the inner lining of the nose and in different organs.

1.3.5 CTC and blood cell classification and its importance

One of the critical research fields involves detecting and classifying circulating tumor cells (CTC) within whole blood cells. CTCs are rare in blood cells, maybe 1 in millions of cells, makes both collection and detection of these cells extremely challenging [79]. Early detection of CTCs has a vital role in early cancer diagnosis and prognosis, provides easy and painless access by a liquid biopsy from blood to identify metastatic cells. Recently, a non-invasive diagnostic method, named a liquid biopsy, of CTCs has emerged as a promising new technique for early cancer diagnosis [80, 81]. It is also expected that the detection and classification of CTCs obtained from the blood may contribute to the diagnosis and treatment of cancer.

1.3.6 Statistical and machine learning approaches in single-cell data.

Advancement in single-cell RNA sequencing has led to the generation of tremendous amounts of data. One of the significant challenges in analyzing such data includes designing efficient and robust machine learning approaches to deal with the noise and sparsity in data. Another challenge is dealing with large-scale single-cell data that include clustering, classification, dimension reduction. Deep learning is making a breakthrough in several areas of bioinformatics, including single-cell RNA-seq data analysis. Auto encoders, long short-term memory (LSTM), and deep generative models such as generative adversarial networks (GANs) are the dominant approaches emerging in single-cell RNA-seq data analysis.

1.4 Scope of the thesis

My goal has been to develop sophisticated statistical and ML-based methods for the precise characterization of single cells in this dissertation. This dissertation addressed three key challenges in this regard, which are illustrated below.

1.4.1 Marker agnostic approaches to detect circulating tumor cells

In the study of single-cell level, a CTC can be defined on at least two criteria. The first is an expression of specific proteins like N-cadherin, *CD44* and epithelial cell adhesion molecule (EpCAM), and the second is an expression of particular genes. The three well-known families of antigen characteristics present

alone or in different combinations within CTC are the epithelial, mesenchymal, and stemness markers [82].

1.4.1.1 Epithelial markers

E-cadherin is one of the essential molecules in cell-cell adhesion in epithelial tissues. It is a part of adherens junctions and one of the hallmarks of epithelial cells. Its function is to enable resistance to force, causing cell detachment and close cooperation with the actin cytoskeleton. It is one of the proteins among others that implicate supporting epithelial tissue architecture. In Epithelial-mesenchymal transition (EMT) progression, E-cadherin plays an important role; when a cell transmits to a mesenchymal state, this protein is decreased [83]. EpCAM is a cell surface glycoprotein that expressed highly in cancerous cells, and low in healthy epithelial cells [84]. If found EpCAM on a cell's surface, it is considered a cell from foreign origin during testing. Cytokeratin (Cks) are markers of normal epithelial differentiation and can be a diagnostic tool to detect different circulating cells of carcinoma. During EMT, their expression is down regulated [85]. Zonula occludens (ZO) are proteins of tight junctions (ZO-1, ZO-2, and ZO-3), involved in epithelial tissue architectural maintenance like E-cadherin. During EMT, one of the earliest modifications is ZO-1 down regulation [86]. Epithelial splicing regulator1 (*ESPR1*), high enrichment of *ESPR1* expression is the indication of the epithelial phenotype, but during EMT, *ESPR1* is down regulated along with *ESPR2*.

1.4.1.2 Mesenchyma markers

N-cadherin expresses in many cells, including mesenchymal cells, was used as a marker of EMT. During EMT, abnormal expression of N-cadherin associated with a dramatic decrease of E-cadherin indicates the mesenchymal character of CTCs switching from E- to N-cadherin [87]. Vimentin is a protein expressed in mesenchymal cells, induces mesenchymal shape of cells, and increases their mobility, considered a marker of EMT [87]. *ZEB1* is an activator of EMT and is implicated in reciprocal regulation with MYB protein and miR-200 family members [88]. It is a DNA-binding transcription factor, a direct transcriptional repressor of E-cadherin. Twist1 is a primary helix-loop-helix transcription factor implicated in tumor development and progression. It is involved in embryogenesis and reactivated in cancers leading to EMT. Fibroblast growth factor receptor 2 (*FGFR2*), an alternative splicing event at the mRNA level, produces either the *FGFR2* IIIb (epithelial) or IIIc isoform (mesenchymal), belongs to a family of trans membrane receptor tyrosine kinases [89]. It causes those cancer cells to display EMT features and migratory behaviour [90].

1.4.1.3 Stemness markers

CD44 is a cell surface glycoprotein which was implicated in cell migration, and metastasis [91]. Aldehyde dehydrogenase-1 (*ALDH1*) is the first biomarker of breast cancer risk, has been shown to identify breast cancer stem cells (CSCs) properties in vivo and in vitro. Gangliosides (GD2, GD3, and GD1a) can be used as markers of CTC stemness. GD2, GD3, GM2, and GD1a were signifi-

cantly increased in CSCs, whereas Fuc-(n) Lc4Cer and Gb3Cer were drastically reduced. In recent, profiling of single-cell has come up with a lot of technology that takes the researcher's attention to explore analysis based on a single cell. Thousand and millions of cells can be profiled at the same time in the same environment. CTC are rarely found in whole blood cells, maybe 1 in millions; recent advancement of single-cell profiling makes it possible to classify CTC to detect it in entire blood cells. This study has many challenges like integration and batch correction of cells from a different environment, a limited number of CTC cells compared to blood cells.

1.4.2 Differential vulnerability of cell types to viral infection

In current COVID-19, infection is spreading with high speed, but due to the advancement of technology in single-cell profiling, explore the area to research fast to prevent infection and diseases. To discover what cells could be involved in COVID-19 transmission, scientists analyzed multiple Human Cell Atlas (HCA) consortium datasets of single-cell RNA sequencing with more than 20 different tissues of non-infected people. These included cells from the lung, eye, kidney, gut, nasal cavity, heart, and liver. The scientists looked for which individual cells expressed both of the two key entry proteins used by the COVID-19 virus to infect our cells. Two markers receptors *ACE2* and *TMPRSS2* are responsible for the entry of virus in cells. In recent, there is no fast and robust technology to test the patient with an infection. Some of the patients are found with no symptoms, and also, this virus has some symptoms shared with other infections. So Its recent advancement of technology in single-cell explores the

area to find particular cells affected by this novel COVID-19 virus. It would also help to develop fast and robust testing phenomena to prevent the spreading of infection.

1.4.3 Noise-free differential expression analysis for robust discovery of marker genes.

In single-cell RNAseq analysis, the major problem is the identification of marker genes. Identification of the Marker gene is of two types. One is cell type's specific marker genes, and another is the marker genes responsible for making differences between two different cell types. Since the last decade, many methods have been developing to find deferentially expressed genes of both types, specifically for the second type. But no one proved a robust way to get deferentially expressed genes specifically in single cells. The single-cell, as compared to bulk cells, has more noise, which makes it more challenging to get gene markers. Some of the methods like SCDE, BPSC, and DESeq2 are used to find deferentially expressed genes for single-cell data, but a significant problem is their long execution time. Among these, SCDE is notably reliable for low false-positive rates but struggles to work on large sample sizes. Recent advanced technologies provide thousands and millions of single cells in one process that explore the study to analyze these many cells together without any bias. But in recent times, methods process to these many cells is challenging due to their time complexity. Thus, in current time, researchers are working on two significant problems: robustness and time complexity.

Chapter 2

Integrative analysis and machine learning based characterization of single circulating tumor cells

A staggering 90% of cancer deaths are attributable to metastases [92]. After detaching from solid tumors, cancer cells travel through the bloodstream to reach distant organs and seed the metastatic tumor's development [93]. As a blood-based biomarker, CTCs offer unabated, real-time insights into tumor growth and therapeutic responses. Despite these promises, owing to their scarcity in the peripheral blood, CTCs are difficult to distinguish and identify [48]. Typically, epithelial tissues are prone to cancer related malignancies. Cancer cells need to develop mesenchymal-like characteristics in order to expand and propagate during metastasis. The transformation of epithelial cancer cells into mesenchymal-like cells is known as epithelial to mesenchymal transition (EMT). Because of the loss of epithelial properties, only a tiny percentage of CTCs are required to express canonical epithelial markers like Epithelial Cell Adhesion Molecule (*Ep-*

CAM). CELLSEARCH® is the only FDA-approved CTC capture tool that uses epithelial surface markers *EpCAM* to detect CTCs in patients' blood [94]. According to controlled experiments involving cell lines [95], the recovery of cells with *EpCAM* expression varies hugely, and certain canonical epithelial markers are down-regulated in CTCs undergoing EMT. Therefore, marker-based enrichment techniques are sub-optimal for the comprehensive charting of heterogeneous CTC subpopulations [96, 97, 98]. Over the past few years, various CTC capture platforms exploiting cancer cells' biophysical characteristics have been developed [99, 100, 101]. *CD45*-based negative enrichment has also been adopted as an alternative strategy. The potential of such antigen-agnostic platforms have not been fully utilized since the chances of immune cell contamination cannot be completely ruled out [99, 100]. The recent advent of scRNA-seq has allowed molecular profiling of single CTCs [42], captured using microfluidic devices [43, 44, 45, 46, 47]. Almost all studies that reported molecular profiles of single CTCs resorted to marker-based bioinformatic annotation of cell types or applied post-capture staining of CTCs using epithelial/cancer-specific molecular markers [102, 43].

2.1 Materials and methods

2.1.1 Description and preprocessing of Datasets

CTCs and peripheral blood mononuclear cells (PBMCs) scRNA-seq data were collected from 14 separate studies in total. [93, 103, 102, 43, 104, 105, 106, 107, 108, 109, 110, 111, 108]. We obtained 558 single CTCs from 10 of the 14 stud-

ies. 6 of these, on the other hand, provided 37665 PBMCs in total. Both blood and CTC transcriptomes are available in two of these studies, with accession numbers GSE67980 and GSE109761, respectively. The CTC data entailed five cancer types: breast, prostate, melanoma, lung, and pancreas. Notably, circulating breast tumor cells in the data were supplied by six different studies. Other single studies represented the remaining cancer types(**Table 2.1**).

Unqiue Studies	Data Provided
GSE51827	CTC
GSE55807	CTC
GSE60407	CTC
GSE67939	CTC
GSE67980	CTC and Blood
GSE74639	CTC
GSE75367	CTC
GSE109761	CTC and Blood
GSE86978	CTC
GSE38495	CTC
EGAS00001002560	Blood
GSE81861	Blood
PBMC 3k	Blood
PBMC 6k	Blood

Table 2.1: Datasets of the studies used in the project with name and types

We identified 15043 genes that were present in all of the samples. First, we discarded the deficient quality cells with less than 10% of the genes having non zero expression. The filtering step retained about 5% (1861) of the input cells. A total of 12335 genes were left after filtering genes with count $\geq$5 in at least ten cells. Our final data contained 12335 expressed genes and 1861 cells, of which 538 were CTCs. At this stage, we standardized the library depths using median normalization [112, 113, 114]. After adding 1 as a pseudo-count, the resulting expression matrix was log-transformed. The subsequent sections discuss the different gene selection strategies and data used for the various downstream studies. While integrating CTC datasets alone, we found 17609 genes

common across all 558 CTCs coming from ten publicly available CTC studies **(Table 2.1)**. We retained CTCs that expressed at least 5% of the 17609 genes. Genes with read count >5 in at least ten CTCs were considered for further analyses. At this stage we were left with an expression matrix consisting of 13600 genes and 554 CTCs.

2.1.2 Construction of epithelial and mesenchymal signatures and E:M Score

We created a panel of 176 well-known epithelial, mesenchymal, and cancer stem cell markers using information from the CellMarker database [115] and current literature. We retained 550 cells that expressed at least 10% of these marker genes. Marker genes having minimum read count >5 in at least 30% of these cells were selected for the subsequent analyses. The resulted matrix consisted of 550 cells and 81 marker genes (16 epithelial, 39 mesenchymal, and 26 cancer stem cell markers, see **(Table 2.2)**. Some of the genes used for all the analysis related to cancer stem cell are *CBX3, HES1, CD58, CHD7, NFIB, SOX4* etc.

Gene	E/M	Behavior in EMT	PMID
AMACR	E	Marker	30198661
CD24	E	Marker	23553902
CDH1	E	Marker	19909494
KRT7	E	Marker	2415537
CAMK2N1	M	Marker	27367674
CTSC	M	Marker	24065739
GADD45B	M	Marker	29629343
IGFBP2	M	Upstream of TGFb	28977895
SVIL	M	Promotes EMT	29954442
TIMP1	M	Induces EMT	24895412

Table 2.2: Some of the genes used for all the analysis related to EMT, E stamds for Epithelial and M stands for Mesenchymal

The generated matrix was median normalised and log-transformed. We cal-

culated a total score for both epithelial and mesenchymal phenotypes for each cell. We used the Z-score transformation on each cell to calculate the score. To create the signature for a specific phenotype, we combined Z-transformed marker expressions using the below formula for each cell.

$$Z_{phenotype} = \frac{\sum_{\forall i \in markers} Z_i}{\sqrt{|markers|}}$$

Here $Z_{phenotype}$ is a comprehensive phenotype-specific score computed over individual Z-transformed marker expressions denoted by Z_i, where $markers$ denotes the set of markers corresponding to the concerned phenotype. We assigned each CTC an E:M scores by computing the ratio between $Z_{phenotypes}$ calculated for epithelial and mesenchymal genes.

2.1.3 Simulation of E-M continuum

We explored the regulatory interactions between epithelial (E) and mesenchymal (M) genes under investigation, as well as their connections to canonical regulators of EMT and the mesenchymal to epithelial transition (MET), such as double-negative feedback loops involving*miR-200, ZEB and GRHL2*. To investigate the mechanistic basis of our data analysis from multiple CTC datasets, we explored the vast literature for a functional implication of the genes that were identified in epithelial and mesenchymal signatures for their roles in EMT and/or MET **(Table 2.2)**. We next constructed a network based on these genes' functional implications in EMT and/or MET, including their effects on the regulatory feedback loops involving miR-200, ZEB and GRHL2 – the fulcrum of epithelial-mesenchymal plasticity.

2.1.4 Classification of cancer and blood transcriptomes

We used various classification models to model the phenotypic identities of CTCs and PBMCs. We used about 3000 cell-type-specific markers identified in the CellMarker database to broaden our feature range, including CDH2, SIGLEC14, GNPDA1, KCNE3, CDH3 etc. [115]. Besides the median normalization, we subjected the data to principal component analysis (PCA) [116] and also applied harmony batch correction method [117]. We used three popular classification techniques - Naive Bayes (NB) [118], Gradient Boosting Machines (GBM) [119] and Random Forest (RF) [120] on the training datasets. We evaluated the model on five different datasets:

1. Clear cell-Polaris CTCs.
2. Hydro-Seq Data, Which uses a novel, hydrodynamic scRNA-seq barcoding technique, for high-throughput CTC capture [101].
3. The leftover PBMCs.
4. A combination of Clearcell-Polaris and randomly sampled leftover 500 PBMC expression profiles.
5. A combination of Hyrdo-seq data and randomly sampled leftover 500 PBMC expression profiles.

2.1.5 Reference component analysis of CTCs and PBMCs

For reference component analysis (RCA), we used the global panels supplied as part of the RCA [121]. RCA [121] uses cell type-specific genes for measuring

the correlation between the tissue types and the input single cells. Due to the low amount of starting RNA, single cell expression data is far noisier than bulk expression data. As a result, tissue types represented by lowly expressed feature genes can give rise to significant noise levels. Therefore, in each global panel, we retained 50% of the tissue types with the highest median expression of the feature genes. RCA [121] analysis provided us with both single cell -tissue correlation heat-map and 2D projection of the individual transcriptomes.

2.1.6 Data and R package availability

The data-set used in the study are available from links mentioned in the **(Table 2.1)**. Single cell sequencing data generated for this work is deposited at GEO with accession number GSE129474. R package is available at GitHub.

2.2 Results

2.2.1 Integration of single cell expression datasets of circulating tumor cells

The majority of our integrative analysis and development of the CTC-immune cell classification system is based on the combined data source **(Figure 2.1)**.

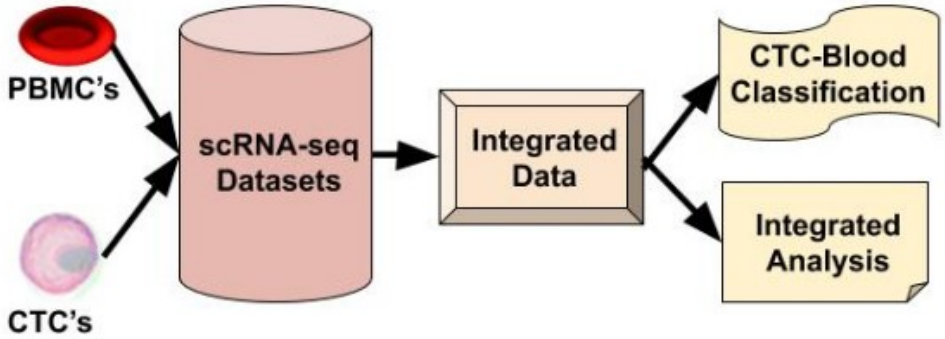

Figure 2.1: Schematic of study

We tracked expression of some of the canonical epithelial (*KRT8, KRT18, EpCAM, CDH1*) and leukocyte markers *(PTPRC, VIM)* to cross-validate the cell type identities. Elevated expression levels of a subset of epithelial markers were observed in a vast majority of the CTCs. Elevated expression levels of a subset of epithelial markers were observed in a vast majority of the CTCs. Significant up-regulation of platelet and fibroblast markers was observed in large fractions of CTCs **(Figure 2.2)**.

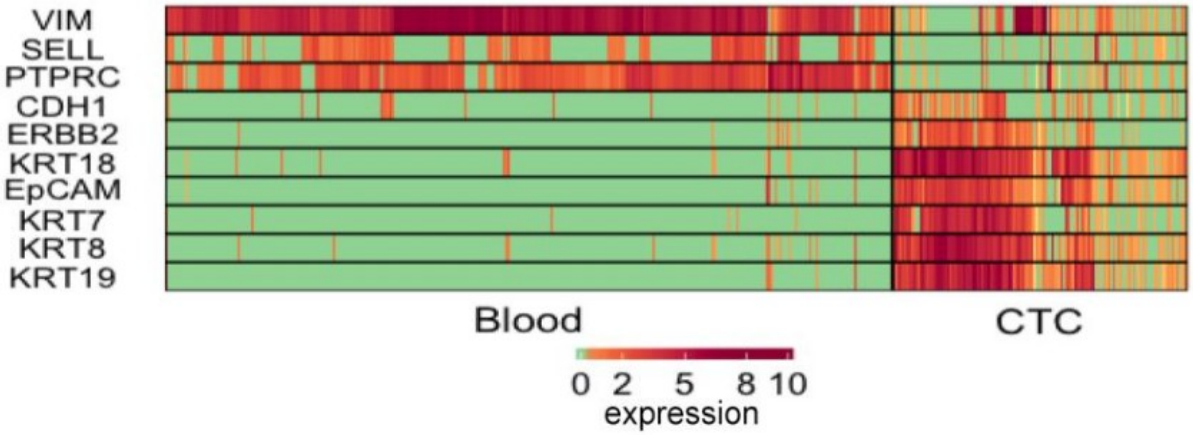

Figure 2.2: Expression of canonical epithelial and immune cell markers in CTCs and the PBMCs under study.

2.2.2 Ubiquity of epithelial-mesenchymal transition in cancer metastasis

EMT and MET have long been postulated to play key roles in cancer metastasis and drug resistance [122]. For each CTC, we computed two scores indicating the strength of epithelial and mesenchymal phenotypes, respectively. We used tens of canonical markers of each of the concerned phenotypes. We detected near-perfect anti-correlation of ($\rho = -0.91$) the phenotypes across CTCs, coming from all cancer types **(Figure 2.3)**.

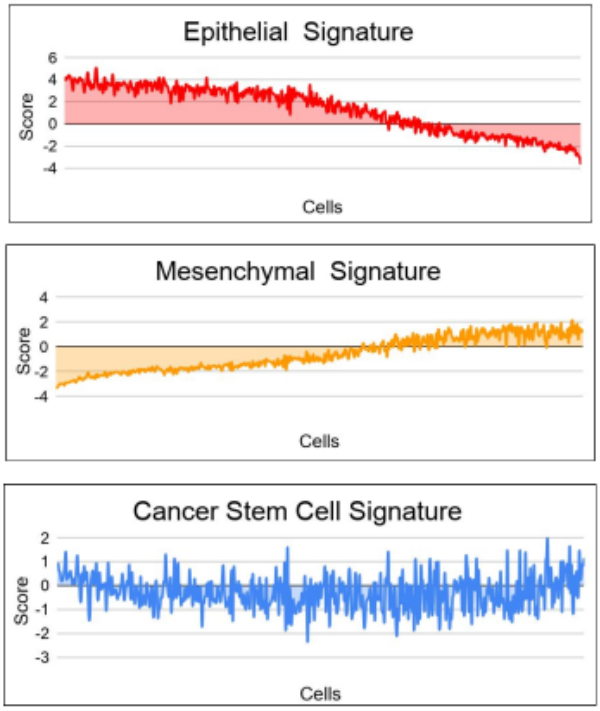

Figure 2.3: Score associated with epithelial, mesenchymal and cancer stem cell signatures across CTCs ordered by E:M score.

Our findings were consistent when we looked at the association between these phenotypes for CTCs in individual studies. Notably, In most of the datasets, CTC transcriptomes were commonly located on an EMT continuum. CTCs were divided into two classes by agglomerative hierarchical clustering, which was primarily based on their approximate binarized identity as epithelial or mesenchymal cells. Despite being on a continuum, CTCs have been found in some studies to form clusters at the epithelial and mesenchymal poles, respectively. Melanocytes come from the neural crest, a multipotent, highly invasive embryonic cell population.The reactivation of the embryonic neural crest program is silenced during normal melanocyte differentiation is thought to be the source of malignant melanoma's high degree of plasticity and aggressiveness [123]. Unlike the CTCs of most cancer types, circulating melanoma cells were found to be clustered exclusively around the mesenchymal pole of the E-M continuum.

Our E:M scores were found to be negatively correlated ($\rho = -0.779$) with EMT score as proposed by Tan and colleagues [124]. One should note that a CTC, enriched with epithelial markers would receive a large positive E:M score, and a large negative EMT score.

2.2.3 Clear patterns observed in expression gradient of immune check-point inhibitor and stemness marker

The activation of cytotoxic T-lymphocytes includes HLA class I (HLA-I) antigens on tumor cells. During natural cancer progression, tumors progressively lose MHC-I expression due to a T-cell mediated immune response, as demonstrated in mouse lines and human cancers selection [125].

The PD-1/PD-L1 pathway, on the other hand, is an adaptive immune resistance mechanism used by tumor cells in response to endogenous anti-tumor immune activity. Tumor cells express PD-L1, which binds to PD-1 receptors on activated T cells, causing cytotoxic T cells to be inhibited [126]. The loss of major histocompatibility complex (MHC) proteins (also known as HLAs) and the activation of PD-L1 together indicate that cytotoxic T cell activities on tumor cells are prevented. Immune checkpoint inhibitors that target the PD-1/PD-L1 pathway have recently emerged as effective cancer therapies [127]. Just a tiny percentage of CTCs in our curated datasets expressed PD-L1. However, there was a consistent anti association between PD-L1 and MHC through studies. One of the datasets with the most PD-L1-activated breast CTCs revealed a connection between PD-L1 and the mesenchymal phenotype.

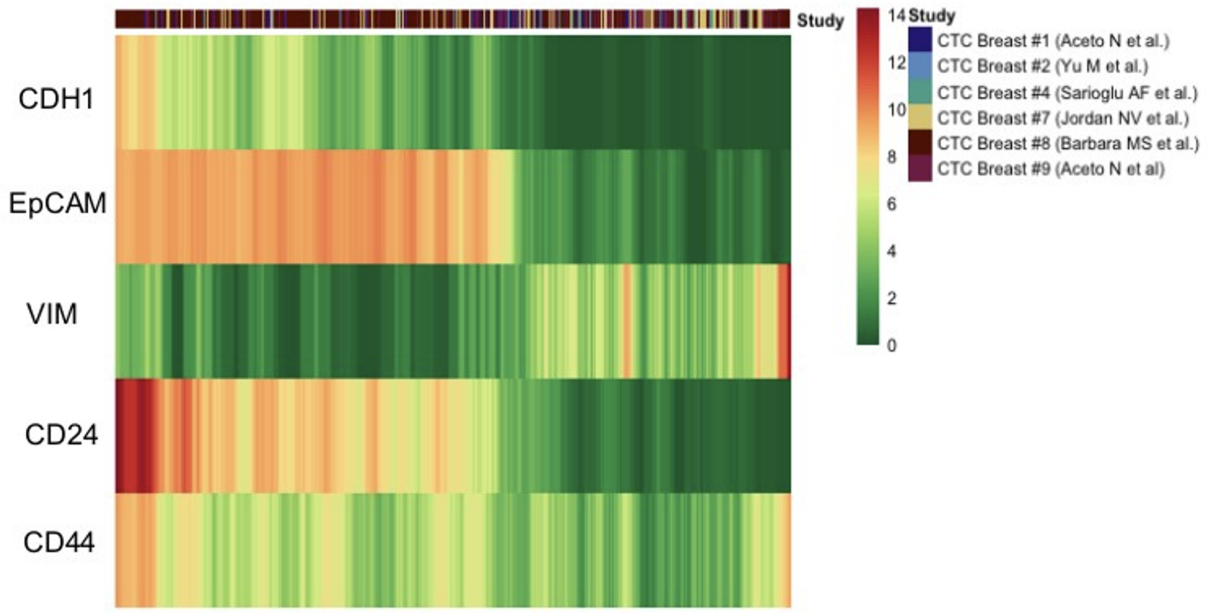

Figure 2.4: The moving average smoothen log expression of well known specific epithelial (CDH1,EpCAM), mesenchymal(VIM) and cancer stem cell markers (CD24, CD44) across breast CTCs,ordered based on the ratio of epithelial and mesenchymal signatures calculated as described in the main methods.

To date, multiple studies have linked EMT to the formation of cancer stem cells (CSCs). In a seminal paper, Mani and colleagues demonstrated the generation of a $CD44^{high}/CD24^{low}$, mammary stem cell-like population due to the induction of EMT. In the mouse, these cells were able to successfully initiate tumors. [128]. $CD44^{high}/CD24^{low}$ CTCs indeed emerge late in the spectrum, following EMT induction. **(Figure 2.4)** This demonstrates how integrative analysis of CTC transcriptomes can help pinpoint stem-like phenotypes, with high tumorogenesis potential.

2.2.4 CTC-PBMC classification system

To train a classifier, we used publicly available single cell expression profiles of human CTCs and PBMCs. Rigorous data preprocessing was performed to

expression datasets culled from various independent studies. Notably, the state of the art batch effect removal method harmony [117] failed to improve the performance of the classification algorithms, compared to a simple median normalisation baseline. We compared the performance of three widely used classifiers - Naïve Bayes [118], Random Forest [120], and Gradient Boosting Machine [119]. We evaluated the model on five different datasets. Overall, the best performing model was GBM with a mean accuracy of ∼93% **(Figure 2.5-B)**. Notably, expression profiles of the CTCs retrieved by the Clearcell-Polaris system were all predicted as CTCs. ∼80% CTCs captured by the recently developed Hydro-Seq [101] (a hydrodynamic RNA-seq barcoding technique, for high-throughput CTC analysis) technique were classified as CTCs.

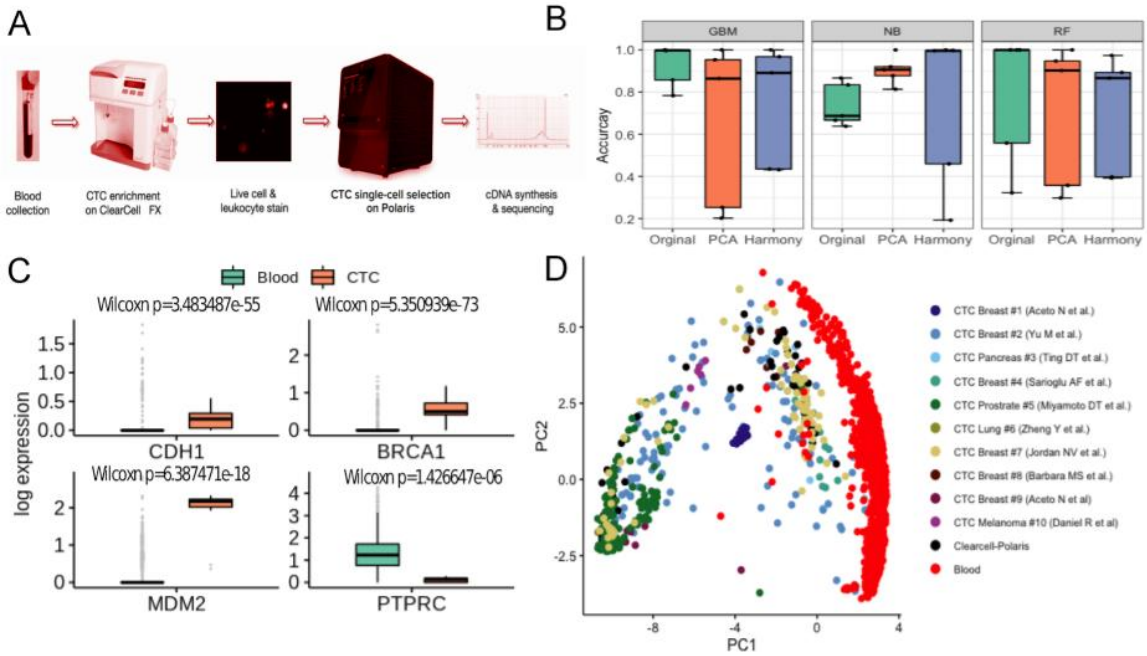

Figure 2.5: Label free detection and characterisation of CTCs. (A) ClearCell-Polaris workflow involving size-based CTC enrichment by ClearCell FX system, followed by single cell selection and CD45/CD31 depletion using Polaris. (B) Performance of various machine learning algorithms in distinguishing between CTCs and PBMCs. Cells in each dataset were tested against a classifier trained on the remaining datasets. Box plots show the prediction accuracy's for different choices of classification algorithms (Naive Bayes or NB, Random Forest or RF, Gradient Boosting Machine or GBM) and normalisation/batch-effect correction methods. (C) Box-plots showing canonical epithelial/breast cancer specific markers, up-regulated in the CTC population compared to the PBMCs. As expected, PTPRC, a pan leukocyte maker shows elevated expression levels in PBMCs as compared to CTCs. (D) Reference Component Analysis (RCA) based 2D projection of CTCs. PBMCs (red) are visibly separated from CTCs. CTCs enriched using the ClearCell-Polaris workflow cluster with CTCs of other types.

2.2.5 Identification of CTCs captured using novel label-free microfluidic workflow

To some extent, existing methods enrich CTCs with contaminating white blood cells (WBCs). Because of this, distinguishing CTCs from immune cells can be difficult.We addressed this challenge by integrating two commercially available microfluidic systems namely Biolidics ClearCell FX System [129] and the Fluidigm Polaris^{TM} system [130] **(Figure 2.5-A)**. In the proposed workflow CTCs are enriched in two steps - size-based enrichment by ClearCell, followed by CD45 (leukocyte marker) and CD31 (endothelial cell marker) based negative selection by Polaris [130]. To validate the workflow and the PBMC-CTC classification system, we processed peripheral blood samples from three HER2-, stage IV breast cancer patients (identified as P3, P4, P5) using a microfluidic interface ensemble.Polaris could retrieve 13, 12 and 32 cells from the blood samples of patients P3, P4, P5 respectively. 15 of these 57 cells passed the filtering criteria. All 15 cells were classified as CTCs.

To determine the carcinogenic origin of the captured cells, we used additional validation criteria. ClearCell Polaris captured cells had elevated expression of breast cancer-specific markers *BRCA1* and *MDM2* (p-value < 0.05) [131] as compared to a group of randomly selected PBMCs. We also detected up-regulation of *CDH1*, a canonical epithelial cell marker. Expression of *CD45* (PTPRC) was considerably low in these cells compared to the PBMC transcriptomes (p-value < 0.05) **(Figure 2.5-C)**. RCA allows noise-free single cell clustering by projecting single cell transcriptomes on reference bulk expression data. We subjected all CTC and PBMC transcriptomes to RCA analy-

sis [121]. ClearCell-Polaris captured CTCs grouped with other CTCs, whereas the PBMCs formed a separate cluster **(Figure 2.5-D)**.

2.3 Discussion

CTCs have been shown to be of prognostic significance in patients with various cancers [93, 111, 102]. We examined the emergence of EMT among CTCs by combining single cell expression profiles from various published studies. To do so, we created the E:M score, which placed CTC transcriptomes on an approximate pseudo-temporal axis of EMT. In principle, our proposed EMT scoring method is similar to Tan and colleagues' method, which focuses on six major cancer types: bladder, colorectal, breast, gastric, ovarian, and lung.Unlike this, we used widely accepted literature curated E and M markers agnostic of the cancer types. It is suspected that a large number of CTCs do not portray the signature of cancer epithelium, largely due to their acquired phenotype that is suitable for migration [111].

We used machine learning techniques to reliably differentiate CTCs from other immune cell types relatively more common. This is accomplished by combining publicly accessible CTC datasets with model training based on machine learning. Our reported ClearCell® Polaris™ workflow, in tandem with the machine learning based CTC-immune cell classification system, for the first time, enables truly unbiased detection of CTCs. We expect a high acceptance rate for our proposed strategy, given the decreasing per-cell cost associated with single cell gene expression screening. Apart from textitEpCAM, a detailed anal-

ysis of CTC transcriptomes enabled us to recognize consistent pan-cancer CTC surface proteins. We examined genes that code for surface proteins that are up-regulated differently in CTCs over in blood cells. Wilcoxon's rank-sum test was used to compare differentially expressed genes in CTCs and blood cells. The P-values obtained using the Benjamin-Hochberg process were subjected to several test corrections (p.adjust function in R). For selecting the differential genes, we used a 0.05 FDR cutoff (DE). DE genes that were expressed in at least 80% of the CTCs were retained. We downloaded a list of the surface proteins from the Cell Surface Protein Atlas (CSPA) database [132] and took intersection with the narrowed set of DE genes. displays the selected markers in the order of the gene-wise fold change values. In addition to *EpCAM*, some of these markers might be useful to broad-base marker dependent capture of CTCs.

Chapter 3

The cellular basis of the loss of smell in 2019-nCoV infected individuals

The recent pandemic due to the uncontrollable spread of novel coronavirus 2019-nCoV has prompted an overwhelming need for diagnostic approaches that can be quickly applied and hence used by people worldwide[133, 134, 135, 136]. In pursuit of this, various workgroups have extensively generated, curated and analyzed virus-centric datasets [137, 138]. Some major efforts include virus isolation from the airway epithelial cells and its genome sequencing [137, 139]. According to comparative genomics, 2019-nCoV is closely related to bat SARS-like coronaviruses (bat-SL-CoVZC45 and bat-SLCoVZXC21) [137]. Notably, the external subdomain of Spike's Receptor-Binding Domain (RBD) of 2019-nCoV shares 40% identity at the amino acid level with other SARS-related coronaviruses [140]. The external subdomain of the RBD, which is responsible for direct interaction with the host receptors, contains the majority of the RBD's amino acid differences [140]. Further, some of the recent reports underscore the role of angiotensin-converting enzyme II (ACE2) as a prominent surface recep-

tor for the cellular entry of 2019-nCoV [141, 142]. Mechanistic insights further revealed the involvement of viral S-protein in assisting strong interaction with the host ACE2 receptor [143]. All of these findings support the hypothesis that ACE2 is involved in viral entry into the host cell. Various groups have traced ACE2 expression in different organs/cell types to determine the tissue or organ level impact of 2019-nCoV [144, 145, 146, 147]. Notably, many of these studies have leveraged single cell sequencing technology to pin-point the cell subpopulation of interest. Collectively, higher ACE2 expression was observed in a range of tissue/cell-types such as epithelial cells of the esophagus, absorptive enterocytes of the intestines, mucosal cells of the oral cavity, proximal tubule cells of the kidney, myocardial cells of the heart, urothelial cells of the bladder, etc, thereby making them potentially vulnerable to the 2019-nCoV infection [144, 145, 146, 147]. These molecular findings are consistent with clinical signs recorded around the globe, with multi-organ failure rising as a significant contributor to infection-related mortality[148]. While the loss of smell and taste has frequently been implicated to 2019-nCoV infection [149, 150, 151], its cellular basis has remained largely unexplored.

The olfactory epithelium includes several distinct cell types, namely horizontal basal cells (HBCs), microvillar cells (MVCs), Bowman's gland cells (BGCs), globular basal cells (GBCs), olfactory ensheathing glia (OEGs), sustentacular cells (SUSs), immature and mature olfactory sensory neurons (iOSNs and mOSNs, respectively) [152]. Among these, olfactory sensory neurons are the key cell types that possess the receptors for odorant detection [152, 153]. In humans, there are at least 400 functionally distinct OSNs [154]. In addition to

these, the olfactory epithelium contains a variety of other cell types that help to maintain tissue architecture and homeostasis. The SUS and MVC subtypes are in control of both metabolic and physical support for the olfactory epithelium [155]. GBCs and HBCs collectively constitute the basal stem cell population and mainly reside near the basal lamina [156]. These cell-types ensure the renewal of the distinct cell types of the olfactory epithelium [157, 158, 159]. Although the aforementioned cell types' mechanisms of action in mediating optimum olfactory function are well understood, further research is required to classify the particular cell types that are susceptible to 2019-nCoV. The susceptibility of olfactory cell types to 2019-nCoV infection was examined in this study. To assess cell-type-specific expression levels of well-known viral-entry host genes, we used a recently published high-throughput single cell expression study. Our meta-analysis revealed that a subset of SUS cells that are enriched for cytoskeleton regulatory proteins are the most vulnerable cell type to the 2019-nCoV infection, followed by a minor population of BGCs, and olfactory stem cells (OSCs: GBCs and HBCs). Aside from humans, we also pinpointed four at-risk mammalian species with high susceptibility to 2019-nCoV infection and the potential of experiencing an infection-mediated loss of olfaction.

3.1 Materials and Methods

3.1.1 Single cell RNA-sequencing analysis

The raw read counts of the single cell RNA sequencing datasets were downloaded from GEO (GSE139522) [160]. We used the widely used Seurat soft-

ware suite [157] for most of our analyses, including cell/gene filtering, clustering, and differential expression analysis[157]. Inbuilt functions NormalizeData(), FindVariableFeatures(), ScaleData(), RunPCA(), DimPlot(), FindNeighbors() and FindClusters() were used for the various standard steps of single cell expression data analysis. Each cluster could be unequivocally mapped to a known cell-type, based on the markers reported in the original study **(Figure 3.1-A, B)** [160].

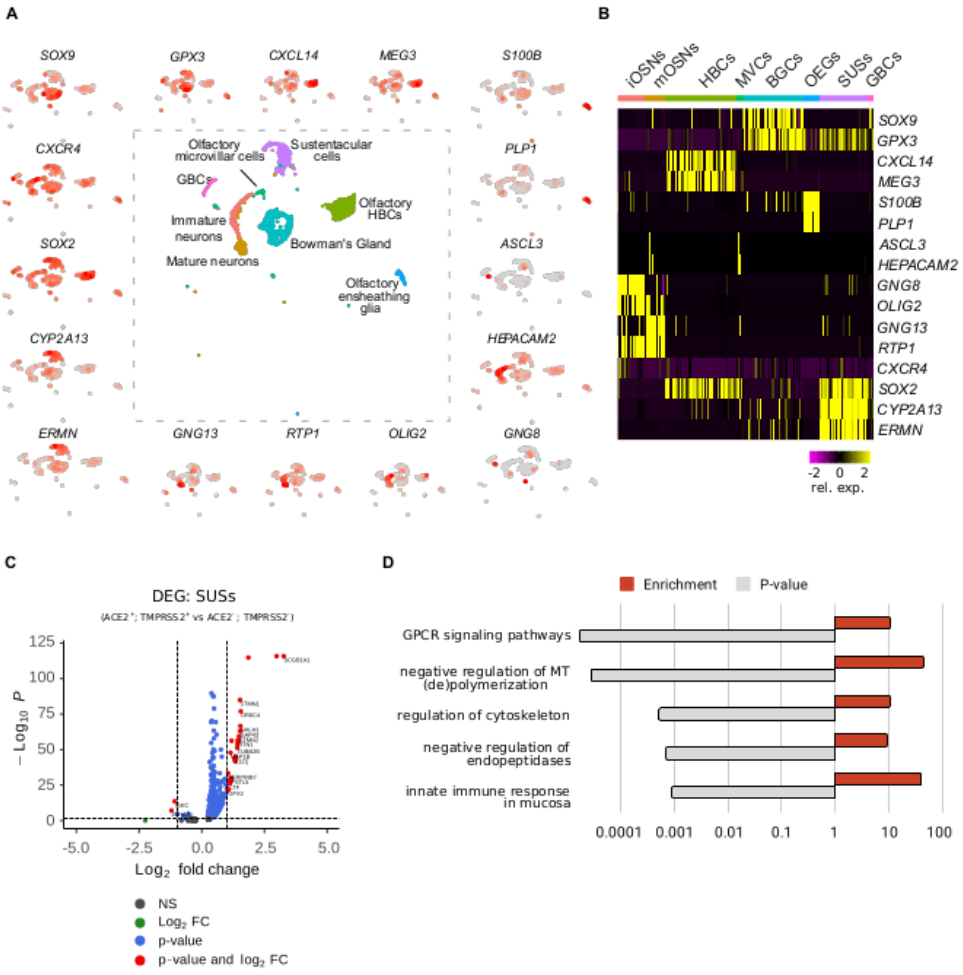

Figure 3.1: Cell-cluster annotation using known bonafide cell-lineage markers. (A) UMAP based embedding of single cell expression profiles represents the relative expression of bonafide markers in the distinct cell types of the human olfactory epithelium. (B) Heatmap depicting the relative enrichment of the marker genes in the indicated cell types of the olfactory epithelium. Scale bar represents the normalized expression values. (C) Volcano plot depicting the significant differentially expressed genes between ACE2+; TMPRSS2+ and ACE2-; TMPRSS2- SUS cells. Y-axis represents the p-value (-log to the base 10) and the x-axis represents fold change (log to the base 2). Significant differentially expressed genes are depicted in red. (D) Bar graph depicting the enrichment and significance of the indicated gene ontologies. Functional enrichment analysis was performed on the significant differentially expressed genes between two subpopulations of SUS cells (ACE2+; TMPRSS2+ vs ACE2-; TMPRSS2-)

The Poisson method, a built-in feature of the Seurat software suite, was used to perform differential gene expression analysis. We constructed average expression vectors for different predefined categories of cells (all eight olfactory cell types, and ACE2 positive cells/TMPRSS2 positive cells/CTSL positive cells/BSG (CD147) positive cells separately) through two biological replicates to ensure reproducibility, as available from the concerned study [160]. The extent of the linear relationship between each of the normalized expression vector pairs was computed using Pearson's correlation coefficient. We also performed a Chi-Square test for comparing the relative proportions of various 2019-nCoV susceptible cell subpopulations between the biological replicates.

3.1.2 Estimating the extent of host-virus protein-protein interactions across cell-types

The host-virus protein-protein interactions were obtained from a recent study [161]. For cell and gene filtering, we used the FilterCells() and FilterGenes() functions respectively from the dropClust pipeline, with the default parameter values [114, 71]. The resulting reduced expression matrix was then median normalised, after adding 1 as a pseudo-count, we log2-transformed the normalised expression estimates. Following that, we retained those N genes for which the corresponding proteins were existed among the host-proteins reported in the study. The second pass of cell-filtering was performed to retain cells that expressed at least 10% of these genes. For each cell, a combined Stouffer's Z score Z was computed as

$$Z = \frac{\sum_{i=1}^{N} Z_i}{\sqrt{N}}$$

, where Z_i denotes the cell-specific Z-score corresponding to i^{th} gene, and N denotes the number of genes common between the expression and the protein-protein interaction data. A Stouffer's score, in this context, reflects the extents of protein-protein interaction in a cell. One-sided Wilcoxon Rank-Sum test was performed to assess the statistical significance of cell type-specific putative enrichment of host-virus protein-protein interactions **(Table 3.1)**.

Cell Types	**iOSNs**	**SuSs**	**HBCs**	**mOSNs**	**BGCs**	**OEGs**	**MVCs**	**GBCs**
iOSNs		1	0.68	1	1	0	1	1
SuSs	0		0	0	0	0	0	0.45
HBCs	0.32	1		1	1	0	1	1
mOSNs	0	1	0		1	0	0.87	1
BGCs	0	1	0	0		0	0	1
OEGs	1	1	1	1	1		1	1
MVCs	0	1	0	0.13	1	0		1
GBCs	0	0.55	0	0	0	0	0	

Table 3.1: Representing the one-sided Wilcoxon Rank-Sum test derived p-values, depicting significance between the cell-type specific distributions of the Stouffer's scores in the indicated conditions.

3.1.3 Analysis of Bulk RNA sequencing dataset

Uniformly processed bulk RNA-sequencing data containing transcriptomic profiles of whole olfactory mucosa from 5 mammalian species i.e. human, monkey, marmoset, mouse, and rat were obtained from a recent publication from Saraiva and colleagues [162]. Log transformed FPKM values were used for plotting the bar charts. The student's t-test was used to calculate the differences in the mean values across the species. A Pvalue $< 0.05, 0.01, 0.001, 0.0001$ is denoted as $*$, $**$, $***$, $****$.

3.1.4 Construction of the phylogenetic tree

The protein sequences of ACE2 of 5 mammalian species were downloaded from the NCBI database to construct the phylogenetic tree.The phylogenetic tree was constructed using an online web server (http://www.phylogeny.fr/) [163]. Protein sequences were supplied to the web server in FASTA format. MUSCLE (version 3.8.31) was used to perform multiple sequence alignment. Gblocks (version 0.91b) was used to refine the alignments, with a minimum block length of 10 and no allowed gap positions. Whelan And Goldman amino acid substitution model was used (available in PhyML 3.1/3.0 aLRT) [164], where the number of substitution rate categories was set to 4. Finally, the tree was rendered using TreeDyn (version 198.3)[165].

3.1.5 Homology modeling and molecular docking

The protein sequences of ACE2 receptors of all species were obtained from the NCBI database (Rat: XP_032746145.1, Mouse: BAB40431.1, Marmoset: XP_-008987241.1, Macaque: NP_001129168.1). For Homology modeling, the human ACE2 was used as the template (PDB ID: 6VW1) [166]. The 3D structures were generated by using Modeller v9.24 [167]. The Discrete Optimised Protein Energy (DOPE) score given by the Modeller, as well as the Ramachandran plots produced by RAMPAGE, were used to evaluate the models' efficiency. Following the refinement of the models, we performed protein-protein interactions between the spike receptor-binding domain (RBD) of 2019-nCoV and the host-specific ACE2 receptors in docking experiments.We used the HAD-

DOCK 2.4 web server for molecular docking experiments [168]. The previously known residues involved in the interaction were used to define the active and passive residues of the proteins. HADDOCK produces ranked clusters of protein-protein complexes based on the HADDOCK score (1.0 Evdw + 0.2 Eelec + 1.0 Edesol + 0.1 EAIR). Notably, the HADDOCK score consists of a combination of empirical (Desolvation, Buried Surface Area) and energy (Electrostatic, Van der Waals) terms. The top resultant complexes were then processed by PRODIGY (PROtein binDIng enerGY prediction) web platform to evaluate their binding energy [169]. As a control, docking of human ACE2 was performed with RBD of SARS-CoV (PDB ID: 2AJF), while using the same docking parameters. Pairwise Mann-Whitney test was performed to compute the statistical significance.

3.1.6 Multiple sequence alignment

The identification of highly conserved residues (i.e. the same residue at the same position in all five species) and partially conserved residues was made possible by aligning the sequences of ACE2 proteins from five different mammalian species (human, rat, mouse, macaque, and marmoset) (i.e. replaced by an amino acid with similar biochemical properties or in other words, a conservative or semiconservative replacement). Multiple sequence alignment was performed using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) [170].

3.2 Result

3.2.1 Divergent expression dynamics of viral-entry genes across olfactory cell subpopulations.

We evaluated the expression of a panel of the known viral-entry transcripts (ACE2, TMPRSS2, BSG/CD147, and CTSL) in 3906 olfactory epithelium originated single cells from the recent report by Durante and colleagues [160], collectively entailing eight distinct olfactory cell types namely HBCs, MVCs, BGCs, GBCs, OEGs, SUSs, iOSNs, and mOSNs. We performed unsupervised clustering of the individual cells using the Seurat software suite [157]. The clusters thus obtained, were unambiguously mapped to specific cell types based on previously known markers **(Figure 3.2-B 3.1-A, B)** [160].

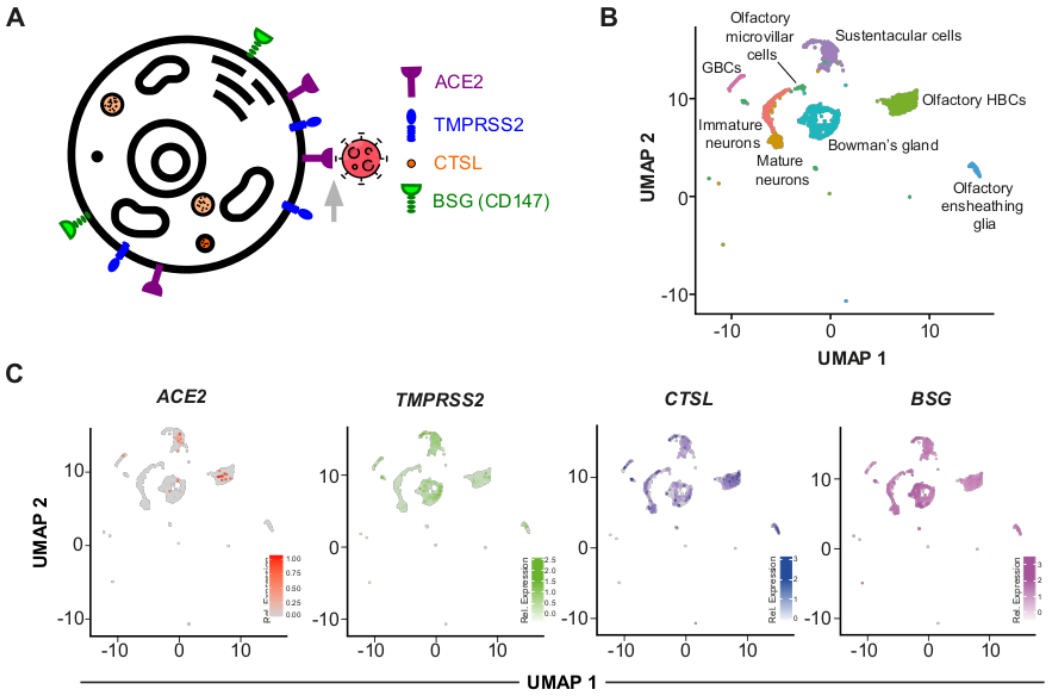

Figure 3.2: Olfactory sensory neurons do not express 2019-nCoV entry genes. (A) Schematic diagram depicting the subcellular localization of the known 2019-nCoV entry host proteins. (B) UMAP based embedding of single cell expression profiles represents the distinct cell types of the olfactory epithelium (C) UMAP based embedding portrays the relative expression of indicated transcripts in the distinct cell types of the human olfactory epithelium.

Each of the eight cell types showed distinct expression profiles for each of the four well-known viral-entry transcripts. We found sparse expression levels of ACE2 in SUSs, BGCs, GBCs, and HBCs, which were mostly restricted to four cell types. Notably, these four cell types collectively constitute less than 1% (32 out of 3906) of the total analyzed cell population. Conversely, transcripts from TMPRSS2, BSG, and CTSL were observed at relatively higher concentrations across all the cell types **(Figure 3.2-C)**.

In addition to binding to host receptors, efficient entry of 2019-nCoV also requires priming of the viral S protein via host proteases [171]. As such, we next investigated the cellular cooccurrence of these essential moieties. Our single-cell co-expression study showed that a subset of SUS cells has a higher infection susceptibility across all combinations**(Figure 3.3-A, B)**.

Notably, due to the lack of direct evidence for the BSG-mediated viral entry into the host cell [147], for further analysis, we focused on the expression dynamics of ACE2 and TMPRSS2. Next, we characterized the phenotypic divergence between ACE2+; TMPRSS2+ and ACE2-; TMPRSS2- subpopulations of the SUS cells by examining the differentially expressed genes (DEGs) **(Figure 3.1-C)**. SUS cells by examining the differentially expressed genes (DEGs) **(Figure 3.1-C)**. Functional enrichment analysis of the significant DEGs (log2FC $\geq$ 1 or $\leq$ -1; FDR $<$ 0.05) revealed the enrichment of the cytoskeleton regulation genes in ACE2+; TMPRSS2+ double-positive SUS cells **(Figure 3.1-D)**. Notably, certain replication machinery components of SARS-CoV, a virus similar in properties to that of 2019-nCoV, utilize microtubule-associated intracellular transport [172].

The most vulnerable olfactory cell types for 2019-nCoV infection were identified using an orthogonal approach based on the host-virus protein interactome. We devised a novel method for overlaying the host-virus protein interactome on cell-type-specific expression signatures to achieve this. Based on the interactome enrichment analysis, we ranked the various olfactory cell-types. In line with our previous analyses, the sustentacular cells were found to be maximally susceptible to viral infection **(Figure 3.3-C, D, Table 3.1)**.

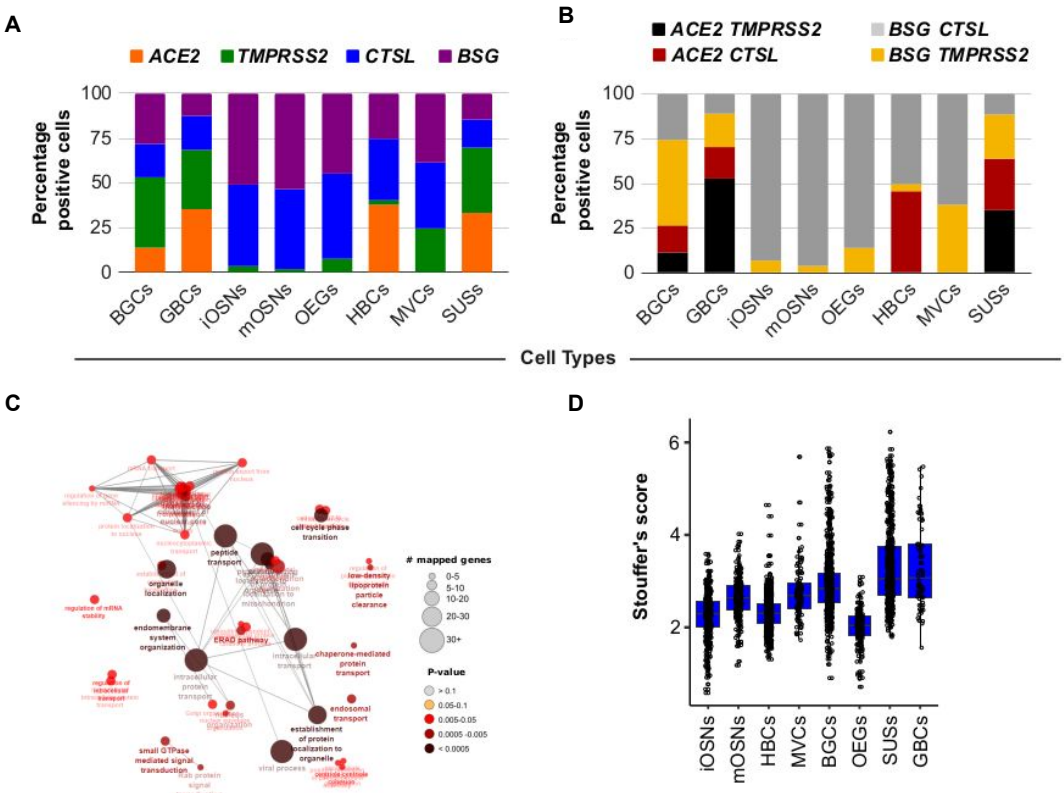

Figure 3.3: Olfactory sensory neurons do not express 2019-nCoV entry genes. (A) Stacked bar graphs representing the relative proportions of cells (percent normalized) expressing the indicated 2019-nCoV-entry associated genes. (B) Stacked bar graph representing the relative proportion of cells (percent normalized) co-expressing the known host-receptor (ACE2 or BSG) and cellular protease (TMPRSS2 or CTSL). (C) Functional enrichment analysis of viral-human protein-protein interactome genes reliably identified in olfactory epithelial cell types. (D) Box plot depicting the Stouffer's score computed based on viral-human protein-protein interaction related genes across indicated cell types of the olfactory epithelium.

Importantly, we conducted a reproducibility study on all single olfactory cell types collected from two biological replicates (patients 2 and 3 of Durante et al. 2020) [160]) to ensure the consistency of our results [160]). For each subpopu-

lation, our analysis revealed highly reproducible expression patterns across the replicates **(Figure 3.4)**.

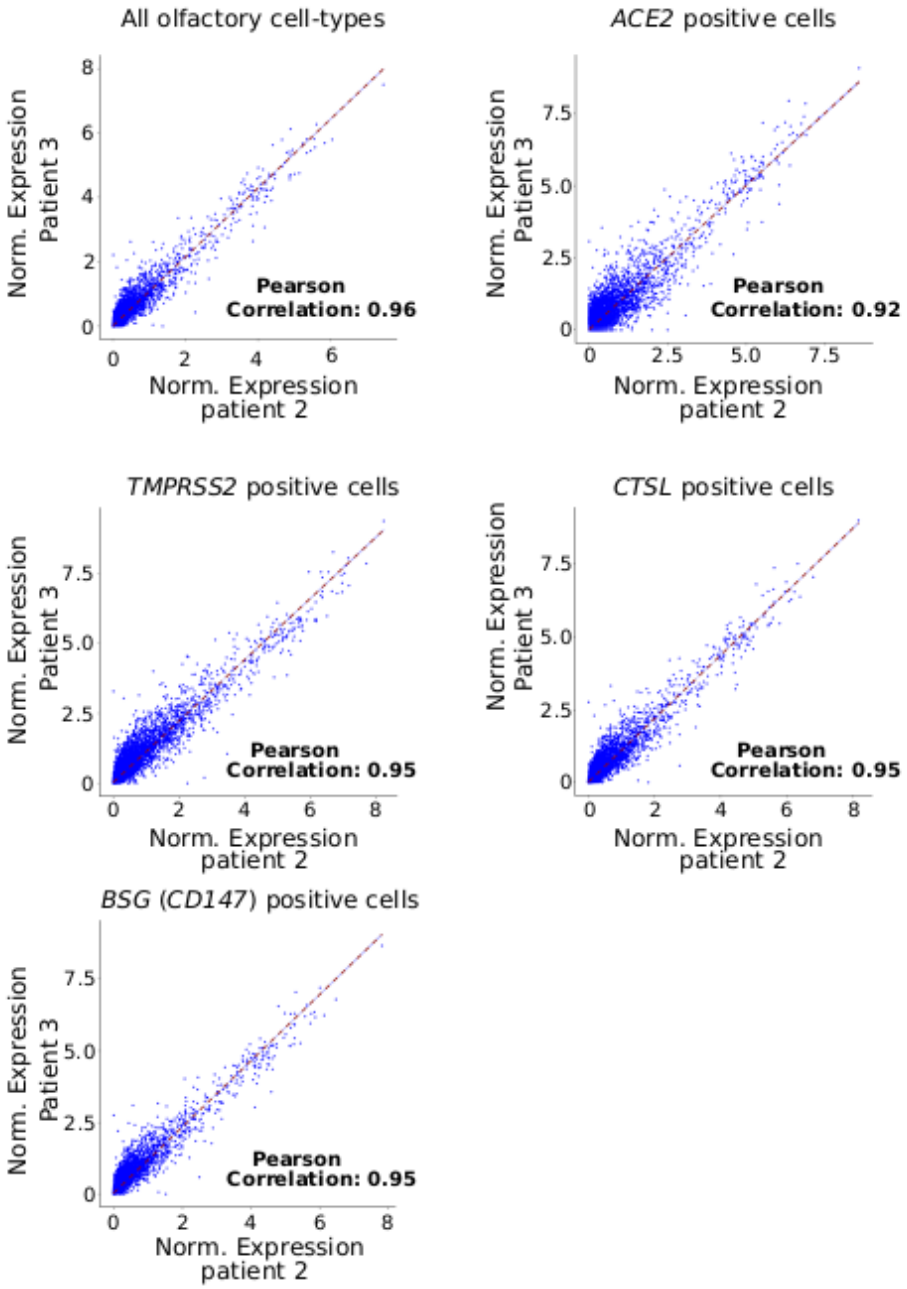

Figure 3.4: Reproducibility analysis of single cell RNA sequencing. Scatter plots depicting the relationship between transcriptomic signatures comprising average expression levels (normalized and log-transformed) of all the filtered genes between two biological replicates corresponding to the indicated subpopulations. All Olfactory cell types correspond to all the eight cell-types, namely HBCs, MVCs, BGCs, GBCs, OEGs, SUSs, iOSNs, and mOSNs.

Moreover, similar results were obtained when the relative proportions of the 2019- nCoV susceptible cells were compared between the biological repli-

cates (ACE2 and TMPRSS2 double-positive cells; Chi-Square value=2.75, p-value=0.43). In summary, while the olfactory sensory neurons largely lack the host-specific proteins essential for the cellular entry of 2019-nCoV, the supporting and the stem cell subpopulations of the olfactory epithelium are potentially highly susceptible to the viral infection.

3.2.2 Comparable expression levels and binding affinity of ACE2 towards viral spike protein across five mammalian species.

The rate of transmission of 2019-nCoV is remarkably higher as compared to the related SARs-CoV [173]. Although it has been speculated that the 2019-nCoV is transmitted to humans from animal sources [173], little is known about the capability of the other mammal species to act as carriers. Notably, in a recent report, monkeys have been confirmed as potential carriers of 2019-nCoV [174]. We wanted to know whether other mammalian organisms are at risk of 2019-nCoV-mediated loss of olfaction. To see if this was true, we looked at the messenger RNA levels of ACE2 and TMPRSS2 transcripts in bulk RNA-Seq profiles of the entire olfactory mucosa of five different mammalian species. Our results suggest a comparable expression of these two viral entry genes (ACE2 and TMPRSS2) among all the species **(Figure 3.5-A, B)**. Next, in order to gain direct evidence of the molecular interactions between the viral S-protein and ACE2, we first modeled and refined the three-dimensional stable protein structure of ACE2 homologs from all the four mammalian species **(Figure 3.6-D)**. To achieve this, we used the recently solved human ACE2 protein structure as a template **(Figure 3.6-A, B, C, D)** Next, we performed molecular docking

between the viral RBD [166] and the modeled ACE2 proteins specific to the individual species (**Figure 3.5-C, D**).

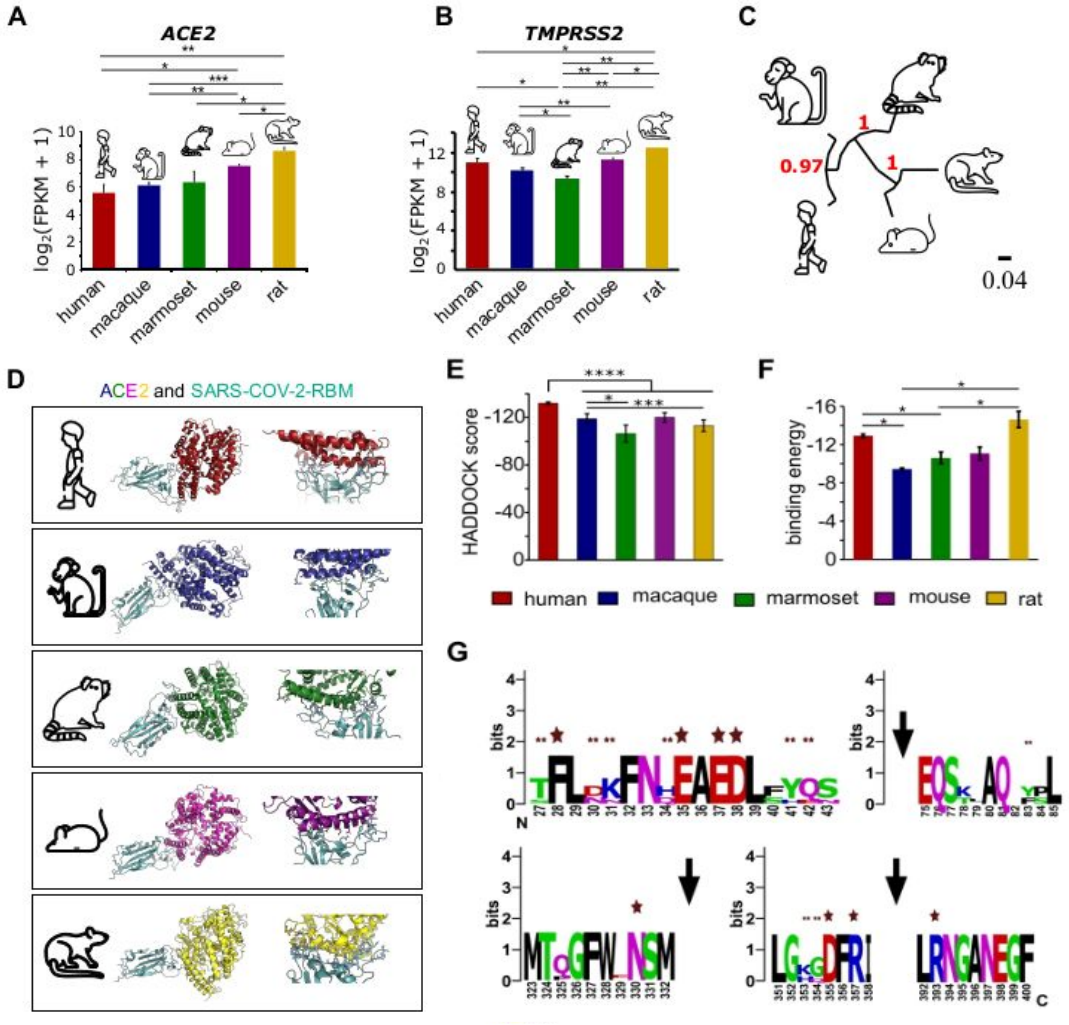

Figure 3.5: Multi-factor analysis involving gene-expression and molecular docking highlights the potential risk of olfactory dysfunction in other mammals. (A) Bar graph depicting the relative abundance of ACE2 in the bulk RNA-sequencing of the whole olfactory mucosa of 5 indicated mammalian species. Bars represent the mean values, the error bars represent the standard deviation, and asterisks represent statistical significance. (B) Bar graph depicting the relative abundance of TMPRSS2 in the bulk RNA-sequencing of the whole olfactory mucosa of 5 indicated mammalian species. The bar represents the mean values, the error bars represent the standard deviation, and asterisks represent statistical significance. (C) Phylogenetic tree depicting the ACE2 sequence similarities between 5 mammalian species. (D) Protein structures depicting the molecular interactions between ACE2 proteins and the RBD domain of 2019-nCoV estimated using computationally-assisted molecular docking. Structure of 2019-nCoV receptor-binding domain (pale cyan) complexed with its receptor ACE2 (distinct color for different species). (E) Bar graph depicting the HADDOCK scores under the indicated conditions. Error bars represent the standard deviation of the estimates, and asterisks represent statistical significance. (F) Bar graph representing the binding energies of the interaction between the 2019-nCoV receptor-binding domain and ACE2 receptor in the indicated species. Error bars represent the standard deviation of the estimates, and asterisks represent statistical significance. (G) Web logo representing the key conserved amino acids of ACE2 of five mammalian species. Single and double asterisks represent highly and partially conserved known interacting residues, respectively

A comparison of the binding parameters showed that all of the tested pairs

had identical binding affinities, implying that these organisms are vulnerable to 2019-nCoV infection (**Figure 3.5-E, F, G**).

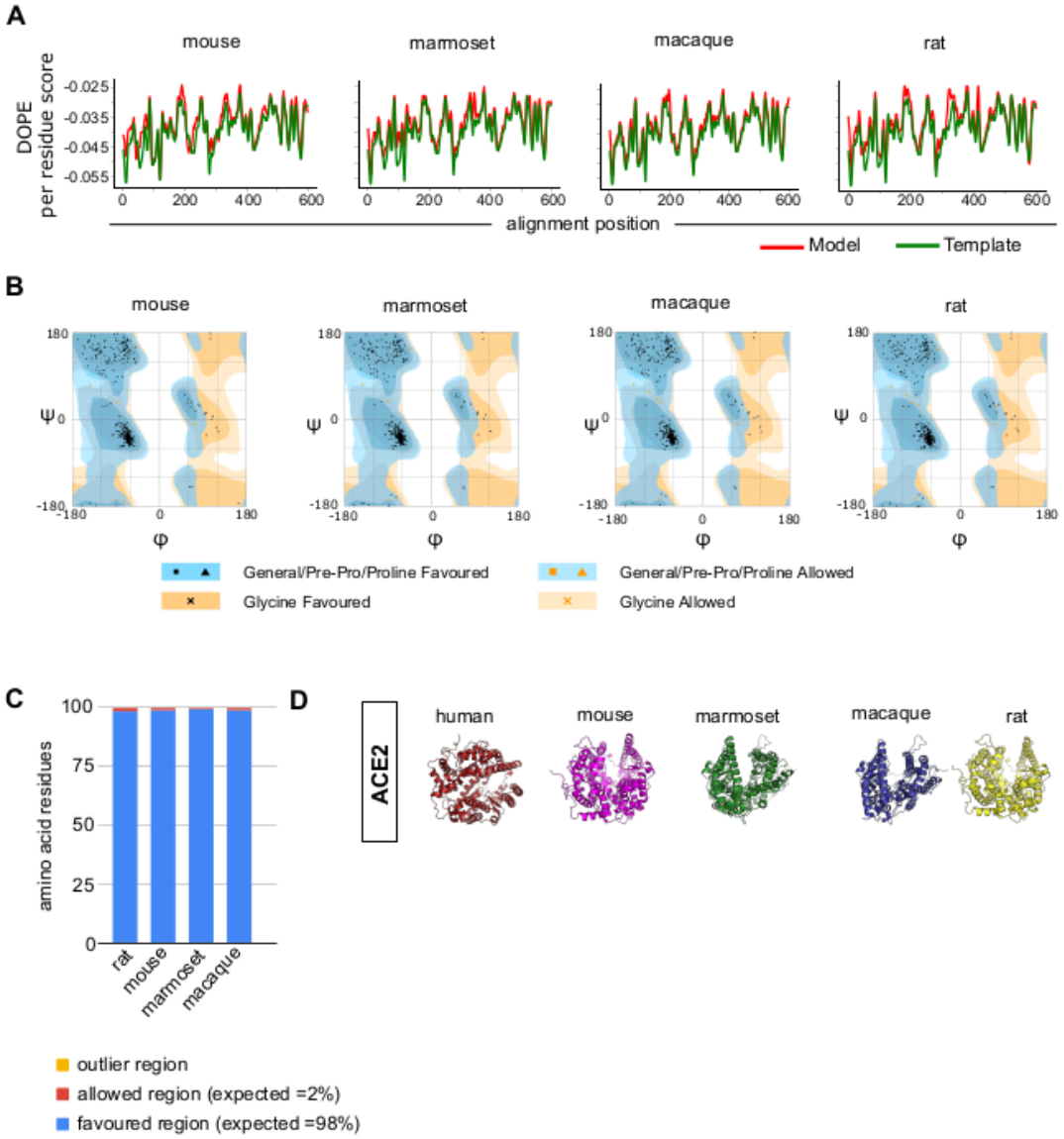

Figure 3.6: Homology modeling based structure prediction of ACE2 proteins from four mammalian species. (A) Line plots depicting the Discrete Optimized Protein Energy (DOPE) scores of the predicted protein structures and the human ACE2 structure (template). Y-axis represents the DOPE score, and the x-axis represents the alignment positions of the amino acid residues. (B) Ramachandran plots depicting the location of the amino acids of the modeled protein structures in the favored, allowed, and outlier regions. (C) Bar graph depicting the percentage of amino acids of the modeled ACE2 structures in the favored, allowed, and outlier regions. (D) Predicted and refined ACE2 structures of the indicated mammalian species.

Importantly, to ensure the robustness of the docking experiments, we compared the HADDOCK scores obtained from docking of human ACE2 with RBD of SARS-CoV and 2019-nCoV. Notably, a recent report experimentally esti-

mated the EC50 (Half maximal effective concentration) values of the aforementioned interactions (human ACE2 and viral RBD domains) [175] and observed preferable binding of 2019-nCoV with that of human ACE2, as compared to SARS-CoV **(Figure 3.7)**. We also observed significantly lower HADDOCK scores in the docking simulations of human ACE2 and 2019-nCoV (pairwise Mann Whitney U test, pvalue < 0.0001), which is in line with the published experimental results [175]. Collectively all these analyses suggest that similar to humans, the olfactory system of other mammals could also be at potential risk of 2019-nCoV infection.

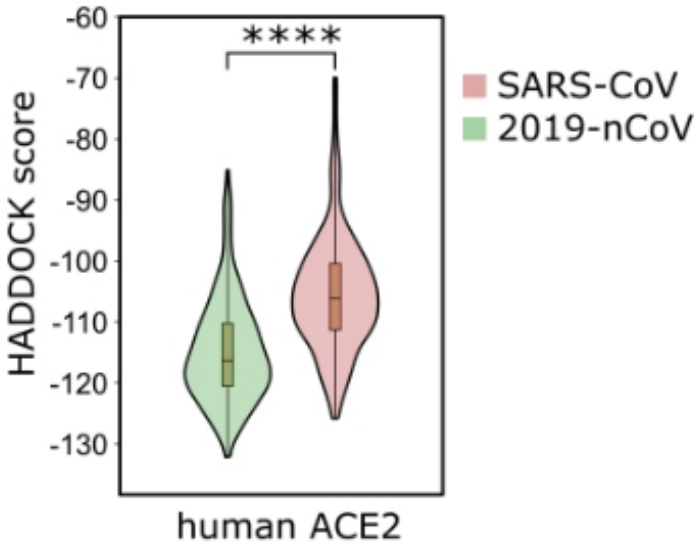

Figure 3.7: Violin plot depicting and comparing the HADDOCK scores obtained from the docking analysis of human ACE2 with the RBD domain of SARS-CoV and 2019-nCoV respectively. Asterisks denote statistical significance.

3.3 Discussion and future directions

A significant bottleneck in fighting the pandemic outbreak of 2019-nCoV, apart from infection-induced multi-organ dysfunction, is the unavailability and inaccessibility of diagnostic methods to the general public worldwide. Despite

the fact that significant efforts have been made to develop 2019-nCoV centric molecular diagnostic kits, the fabrication, production, mass-distribution, and acceptance of these kits is likely to take time.ng. Recently, multiple clinical studies have reported the abrupt loss of smell and taste in a large number of 2019-nCoV infected individuals [149, 150, 151], thereby, collectively reinforcing its potential application as the first line of diagnostics in the patients exhibiting other 2019-nCoV-related hallmark symptoms. In pursuit of this, The Global Consortium of Chemosensory Researchers (GCCR) has initiated scientific investigation into the potential links between respiratory disease and its impact on smell and taste. Other reports have also suggested that the sudden loss of olfaction is the first symptom in reported 2019-nCoV infected patients. [149, 150, 151]. Additionally, the symptom survey of 2019-nCoV infected patients also revealed the loss of smell as a stronger predictor of positive diagnosis than the self-reported fever [149]. Our research examines the olfactory-epithelium-specific cell-types based on the expression levels of host-specific viral-entry moieties as well as the burden of host-virus protein-protein interactions in order to highlight the possible cellular basis of olfactory loss. Our findings indicate that the infected patients' loss of smell may not be due to a direct impairment of the olfactory sensory neurons. Instead, SUS cells, BGCs, and olfactory stem cells (OSCs: HBCs and GBCs) exhibit the molecular make-up that makes the cells susceptible to viral infection **(Figure 3.8)**. A consensus approach involving gene expression as well as the host-pathogen protein interactome led to this conclusion. Importantly, since all of our results are largely supported by in silico analysis, the single cell transcriptomics assay's limita-

tions, such as sampling bias and high dropout rates, cannot be ignored [176]. While the identification of the 2019-nCoV infection susceptible cell types of the olfactory epithelium is characterized based on the handful of known viral entry moieties of the host cells, one cannot rule out the possible involvement of currently uncharacterized host cell surface receptors or proteases which may facilitate the viral entry into the host cells.

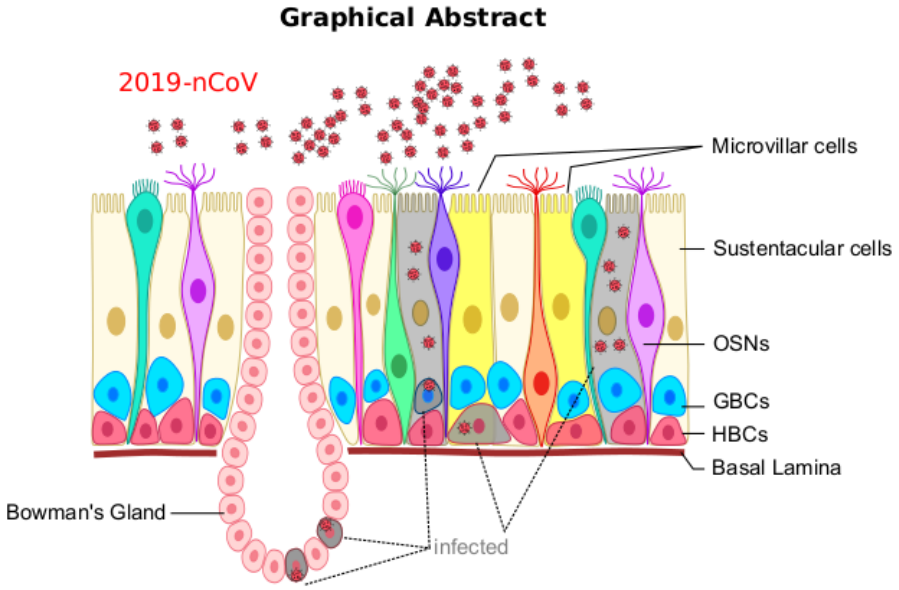

Figure 3.8: Graphical representation of the key findings.

The SUS cells' direct sensory roles are largely elusive, but they are known to provide metabolic and physical support to the olfactory epithelium [177]. Particularly the SUS cells are known to be involved in secretion [178], endocytosis [179], and Cytochrome P-450-mediated detoxification [180]. Moreover, as glia-like cells, they impart critical functionality related to phagocytosis of dead cells [181] and regulation of the ionic exchange with the extracellular regions [182, 183]. Hegg et. al. [184] identified a key function of SUSs in establishing communication between neurons, basal cells, and SUSs themselves. The au-

thors identified that the activation of G-protein coupled receptors, particularly the P2Y purinergic receptor and the muscarinic acetylcholine receptor induce calcium oscillations in sustentacular cells. They further provided mechanistic insights by using pharmacological interventions and showed the involvement of phospholipase C (PLC) pathways in triggering the calcium increase. Notably, in addition to SUSs, our results also highlight the viral infection susceptibility in minor subpopulations of BGCs and OSCs. BGCs play a vital role in maintaining the optimal functionality of the olfactory system. First, they make a number of olfactory binding proteins that help odorants reach olfactory receptor cells. Second, they secrete mucus that prevents the olfactory epithelium from drying out, which helps mOSNs recognise odours indirectly. [177]. Similarly, while OSCs are not known to have any direct role in odorant detection, they play a crucial role in the regeneration of the olfactory epithelium upon lesions [158, 185]. Notably, injury models involving the direct loss of SUSs have been shown to activate the HBCs which in turn proliferate and replenish the lost cells, thereby reconstituting the olfactory epithelium homeostasis [185]. Mechanistically, loss of Notch signaling pathways between SUSs and HBCs leads to the breakdown of mitotic dormancy of HBCs by downregulating tumor protein p63 [186]. Given these important functional roles, we hypothesise that the apparent lack of smell in SUSs, BGCs, and OSCs may be due to viral load. Because of their apical location in the olfactory epithelium, 2019-nCoV can infect SUS cells first, causing a partial or complete breakdown of the olfactory architecture and olfaction loss. Furthermore, due to a breakdown in the repair response as a result of the subsequent infection in the OSCs, the severity of this phenotype

is exaggerated. In addition to this, our analysis on four additional mammalian species also suggests that the 2019-nCoV-mediated loss of smell phenotypes is not restricted to humans, but may also impact other mammalian species. Our study collectively provides the first line of evidence that a subpopulation of olfactory cells is potentially equipped with host-specific viral-entry moieties that the virus can exploit for its entry.

Chapter 4

ROSeq: Modeling expression ranks for noise-tolerant differential expression analysis of scRNA-Seq data

Single cell RNA-Sequencing (scRNA-seq) has significantly accelerated the characterization of molecular heterogeneity in healthy and diseased tissue samples over the last few years[39]. The declining cost of library preparation and sequencing have fostered the adoption of single cell transcriptomics as a routine assay in studies arising from diverse domains, including stem cell research, oncology, and developmental biology [187, 188]. The field of single cell transcriptomics is plagued by a slew of data quality problems, owing primarily to a scarcity of starting RNA content. Single cell gene expression modelling is hampered by high levels of noise and technological bias, which makes it difficult to reach scientifically valid conclusions regarding cell-type-specific gene expression patterns [189]. A number of parametric and nonparametric methods have already been proposed for modeling single cell expression data and find-

ing differentially expressed genes (DEGs). SCDE [36], MAST [190] and BPSC [191] are notable among these. SCDE and MAST model gene expression using well-known probability density functions and mixture models involving some of those. BPSC, on the other hand, handles single cell expression bimodality by employing a Beta-Poisson mixture. Different from these, we conjectured that considering expression ranks instead of absolute expression estimates would make a model less susceptible to the noise and the technical bias, as commonly observed in single cell data. To realize the same, we employed Discrete Generalized Beta Distribution (DGBD) [192] to model the distribution of expression ranks instead of the raw count. The seminal work inspired the consideration of rank-ordering distribution by Martinez-Mekler and colleagues. They demonstrated the universal applicability of the same in linking frequency estimates and their ranks [192]. We developed ROSeq, a Wald-type test to determine differential expression from scRNA-seq data.

4.1 Materials and methods

4.1.1 Description and preprocessing of datasets

For the various analyses, we used four publicly available single cell RNA sequencing (scRNA-seq) datasets. The datasets are named after the surnames of the first authors for better readability. Among these, the Trapnell data contains scRNA-seq profiles of 77/99 primary myoblasts sampled before/24 hours after differentiation [193]. Tung data consists of single cell transcriptomes of induced pluripotent stem cells (iPSCs) generated from three different individuals,

marked as NA19098, NA19101, and NA19239, respectively [1]. For each of the three individuals, a total of 288 cells were profiled. For each condition/individual in the Trapnell and Tung datasets, three bulk RNA-seq replicates were available from the respective studies. Chu dataset consists of undifferentiated H1 (n = 212) and H9 (n = 162) human ES cells and neuronal progenitor cells or NPC (n = 173), with a total of 9 matched bulk replicates (H1=4, H9=3 and NPCs=2) [2]. To diversify our experiments, we used Zheng dataset containing 3258 single cell transcriptomes of Jurkat cells, processed using the GemCode technology [3]. Single cell datasets with a good number of matched bulk replicates are scarce, constraining the validation of DEG callers. We produced scRNA-seq data and matching bulk replicates of human foreskin BJ fibroblast (150 single cells and 3 bulk replicates) and K562 (352 single cells and 4 bulk replicates) to facilitate extensive bench marking. The following section describes the details pertaining to the laboratory methodologies. Bioinformatic processing including read alignment and expression quantification mirrors our previous report [194]. Collectively the BJ/K562 scRNA-seq data is referred to as the Gupta dataset.

We first filtered out cells with less than 2000 detected (non zero read count) genes for each dataset. Gene filtering followed the cell filtering step. We retained the genes having read count ≥ 3 in at least 3 cells [194]. Next, the pruned count matrix was subjected to different normalization techniques depending on the target differential expression method. For Wilcoxon's rank-sum test, BPSC, and MAST, count per million (CPM) normalization was used, following the recommendation by Soneson and Robinson [195]. SCDE and DESeq2 [196] were supplied with the processed raw count data as input. For ROSeq, we first

subjected the processed raw count matrices to the trimmed mean of M-values (TMM) normalization [18], followed by Voom transformation [197].

4.1.2 Mapping expression estimates to ranks

ROSeq accepts normalised read count data as input for gene expression modelling. ROSeq first defines a gene's range by pooling the normalised expression estimates for all cell groups for determining the minimum and maximum values for each gene. Next, the range is split into $k \times \sigma$ sized bins, where k is a scalar with a default value of 0.05, and σ is the standard deviation of the pooled expression estimates across the cell-groups. Each of these bins is assigned a rank, based on the sequential order of its expression range. At the level of a cell-group, this leads to mapping of bin-wise cell frequencies to ranks, such that the bin with the highest cellular frequency is assigned the least rank (i.e., 1). The Discrete Generalized Beta Distribution (DGBD) is used as a probability mass function to express a normalized bin-wise cell-frequency y_r as a function of its corresponding rank r using two real parameters a and b. In other words, the DGBD formulation can be thought of as a discrete distribution of the rank-frequencies. If N be the total number of bins for a given gene, then the DGBD specifies the probability p_r for the r-th rank to have a (relative) size of y_r, which can be expressed as

$$p_r = A\frac{(N+1-r)^b}{r^a}, \quad r = 1, \ldots, N, \tag{4.1}$$

where A is the normalizing constant ensuring $\sum_r p_r = 1$. Note that the sum of the normalized frequencies also equals one ($\sum_r y_r = 1$).

4.1.3 Estimation of the DGBD parameters

For a given gene and a specific cell-group, the best-fitting parameter values ($\widehat{a}$, $\widehat{b}$) are determined by maximizing, with respect to (a, b), the Log-Likelihood corresponding to the model given by Equation 4.1. Considering the discrete probability distribution structure of the DGBD formulation of (relative) rank-sizes, the resulting likelihood function is given by

$$\mathbf{L} = \prod_{r=1}^{N} p_r^{y_r} = A \prod_{r=1}^{N} \frac{(N+1-r)^{by_r}}{r^{ay_r}}.$$

Now, taking logarithm, the required Log-Likelihood function, **logL**, can be computed as (in Equation 4.2)

$$\mathbf{log(L)} = \log(A \prod_{r=1}^{N} \frac{(N+1-r)^{by_r}}{r^{ay_r}}).$$

$$\mathbf{logL}(a, b) = -a \times \sum_{r=1}^{r=N} y_r \log(r) + \tag{4.2}$$
$$b \times \sum_{r=1}^{r=N} y_r \log(N+1-r) + \log(A).$$

The resulting estimates $(\widehat{a}, \widehat{b})$ correspond to the DGBD under which the observed data is most likely to be generated. Such maximum likelihood estimates (MLE) are the most efficient (least standard error) and enjoy several optimum

properties on large sample sizes [198].

To test differential expression of a gene between two cell-groups, based on the above MLEs $(\widehat{a}, \widehat{b})$, we additionally need estimates of their standard errors (equivalently their variance). From the theory of maximum likelihood [199], the asymptotic variance of $(\widehat{a}, \widehat{b})$ is given by the inverse of the associated Fisher information matrix $I(a, b)$, which can be consistently estimated by $I(\widehat{a}, \widehat{b})$. For the log-likelihood function of the DGBD model given in Equation 4.2, the form of the Fisher information matrix I may be simplified in a more succinct form as follows.

$$
\begin{aligned}
I(a,b) &= -\begin{bmatrix} \dfrac{\partial^2 \text{logL}}{\partial a^2} & \dfrac{\partial^2 \text{logL}}{\partial a \partial b} \\ \\ \dfrac{\partial^2 \text{logL}}{\partial b \partial a} & \dfrac{\partial^2 \text{logL}}{\partial b^2} \end{bmatrix} \\
&= A^2 \left(\sum_{r=1}^{N} y_r \right) \begin{bmatrix} u_{2,0}u_{0,0} - u_{1,0}u_{1,0} & u_{1,0}u_{0,1} - u_{1,1}u_{0,0} \\ \\ u_{1,0}u_{0,1} - u_{1,1}u_{0,0} & u_{0,2}u_{0,0} - u_{0,1}u_{0,1} \end{bmatrix}, \qquad (4.3) \\
&= A^2 \left(\sum_{r=1}^{N} y_r \right) \begin{bmatrix} u_{2,0}u_{0,0} - u_{1,0}^2 & u_{1,0}u_{0,1} - u_{1,1}u_{0,0} \\ \\ u_{1,0}u_{0,1} - u_{1,1}u_{0,0} & u_{0,2}u_{0,0} - u_{0,1}^2 \end{bmatrix},
\end{aligned}
$$

where, for each $i, j = 0, 1, 2$, we define

$$u_{i,j} = \sum_{r=1}^{N} \frac{(N+1-r)^b}{r^a} (\log r)^i (\log(N+1-r))^j.$$

Note that, $u_{0,0} = 1/A$. For our DGBD model with likelihood function given by 4.2, we have

$$\frac{\partial^2 \text{logL}}{\partial a^2} = \left(\sum_{r=1}^{r=N} y_r \right) \frac{\partial^2 \log(A)}{\partial a^2}$$

$$\frac{\partial^2 \text{logL}}{\partial b^2} = \left(\sum_{r=1}^{r=N} y_r \right) \frac{\partial^2 \log(A)}{\partial b^2} \tag{4.4}$$

$$\frac{\partial^2 \text{logL}}{\partial a \partial b} = \left(\sum_{r=1}^{r=N} y_r \right) \frac{\partial^2 \log(A)}{\partial a \partial b}$$

So in order to evaluate the above mentioned double derivatives, the first order

derivative $\frac{\partial \log \mathrm{A}}{\partial a}$ and $\frac{\partial \log \mathrm{A}}{\partial b}$ are determined as follows:

$$\begin{aligned}
\log \mathrm{A} &= -\log\left(\sum_{r=1}^{r=N}\frac{(N+1-r)^b}{r^a}\right) \\
\frac{\partial \log \mathrm{A}}{\partial a} &= \frac{1}{\left(\sum_{r=1}^{r=N}\frac{(N+1-r)^b}{r^a}\right)} \times \sum_{r=1}^{r=N}\frac{(N+1-r)^b \log r}{r^a} \\
\frac{\partial \log \mathrm{A}}{\partial b} &= \frac{-1}{\left(\sum_{r=1}^{r=N}\frac{(N+1-r)^b}{r^a}\right)} \times \sum_{r=1}^{r=N}\frac{(N+1-r)^b \log(N+1-r)}{r^a}
\end{aligned} \tag{4.5}$$

Re-writing Equation 4.5 in a more succinct form in the Equation 4.6 below, we get

$$\frac{\partial \log \mathrm{A}}{\partial a} = \frac{u_{1,0}}{u_{0,0}} \quad \text{and} \quad \frac{\partial \log \mathrm{A}}{\partial b} = -\frac{u_{0,1}}{u_{0,0}} \tag{4.6}$$

In a compact form, these can be written more generally, for any $i, j = 0, 1$, as

$$\frac{\partial u_{i,j}}{\partial a} = -u_{i+1,j}, \qquad \frac{\partial u_{i,j}}{\partial b} = u_{i,j+1}. \tag{4.7}$$

Substituting the above expressions in the formula for Fisher information matrix in 4.3, we get its simplified form for computation within our ROSeq.

Evaluating the partial derivatives of $u_{1,0}$, $u_{0,0}$ and $u_{0,1}$ with respect to a and

b, in the Equation 4.8:

$$\begin{aligned}
\frac{\partial u_{1,0}}{\partial a} &= -\sum_{r=1}^{r=N} \frac{(N+1-r)^b \left(\log r\right)^2}{r^a} \\
\frac{\partial u_{1,0}}{\partial b} &= \sum_{r=1}^{r=N} \frac{(N+1-r)^b \left[\log r\right] \left[\log\left(N+1-r\right)\right]}{r^a} \\
\frac{\partial u_{0,0}}{\partial a} &= -\sum_{r=1}^{r=N} \frac{(N+1-r)^b \log r}{r^a} \\
\frac{\partial u_{0,0}}{\partial b} &= \sum_{r=1}^{r=N} \frac{(N+1-r)^b \log\left(N+1-r\right)}{r^a} \\
\frac{\partial u_{0,1}}{\partial a} &= \sum_{r=1}^{r=N} \frac{(N+1-r)^b \left[\log r\right] \left[\log\left(N+1-r\right)\right]}{r^a} \\
\frac{\partial u_{0,1}}{\partial b} &= -\sum_{r=1}^{r=N} \frac{(N+1-r)^b \left[\log\left(N+1-r\right)\right]^2}{r^a}
\end{aligned} \tag{4.8}$$

4.1.4 Testing for differential expression: Two-sample Wald Test

Further, in order to statistically test if a gene is differentially expressed between two sub-populations, ROSeq uses the (asymptotically) optimum two-sample Wald test based on the MLE of the parameters and their asymptotic variances, given by the inverse of the Fisher information matrix.

Let us assume that the DGBD parameters corresponding to the contrasting cell-groups 1 & 2 are denoted by (a_1, b_1) and (a_2, b_2), respectively, and their MLEs based on the available normalized expression data are given by $(\widehat{a}_1, \widehat{b}_1)$

and $(\widehat{a}_2, \widehat{b}_2)$ with the respective number of bins being m and n. We can estimate the asymptotic variance matrices for these MLEs, using Equation 4.3, as $\widehat{V}_1 = I(\widehat{a}_1, \widehat{b}_1)^{-1}$ and $\widehat{V}_2 = I(\widehat{a}_2, \widehat{b}_2)^{-1}$, respectively. Under our the DGBD model, the desired testing for differential gene expressions is equivalent to the test for the null hypothesis $H_0 : a_1 = a_2, b_1 = b_2$ against the omnibus alternative. The Wald test statistic T for testing H_0 can be written as follows:

$$T = \left(\frac{mn}{m+n}\right) \begin{bmatrix} \widehat{a}_1 - \widehat{a}_2 \\ \widehat{b}_1 - \widehat{b}_2 \end{bmatrix}^T \left(w\widehat{V}_1 + (1-w)\,\widehat{V}_2\right)^{-1} \begin{bmatrix} \widehat{a}_1 - \widehat{a}_2 \\ \widehat{b}_1 - \widehat{b}_2 \end{bmatrix},$$

where $w = \frac{n}{m+n}$. If the null hypothesis H_0 is correct, i.e., the genes in the two sub-populations are not differentially expressed, the above test statistics T asymptotically follows a central chi-square distribution χ^2_2 with two degrees of freedom. Therefore, we conclude that the genes are differentially expressed (i.e., reject H_0) at 95% level of significance, if the observed value of the test statistics T exceeds the 95% quantile of the χ^2_2 distribution (which is approximately 6). The corresponding *P*-value is given by the probability that a χ^2_2 random variable exceeds the observed value of T.

4.1.5 Benchmarking of single cell DEG calls

We used matched bulk RNA-seq data from the same studies to benchmark single cell DEG calls. DEG calls were made using DESeq based on bulk RNA-seq data [200]. DESeq's standard pipeline uses median of ratios method of normalization. DEGs were selected at an FDR cutoff of 0.05. To ensure the trustworthiness of the bulk based DEG calls, we imposed a strict fold change criterion

(i.e., $\log_2$ fold change cutoff of 3) as recommended elsewhere [201, 202].

4.1.6 Dropout induction in real scRNA-seq data

Given a count matrix, to simulate dropouts, we computed $E_g = \log_2(R_g + 1)$, where R_g denotes average read count of g across the input transcriptomes. We also computed log-odd of the dropout probability D_g for a gene g, where $D_g = \log \frac{p_g}{1-p_g}$. Here p_g denotes the observed probability of dropouts for g. We modeled D_g w.r.t. E_g using linear regression as indicated below.

$$D_g = \alpha + \beta E_g, \tag{4.9}$$

which simply describes a line with slope β and y-intercept α. As dropout rate increases with decrease in expression, one would expect $\beta < 0$. We confirmed this by visualising the relationship between D'_g and E_g. Of note, Splatter, a popular dropout induction method makes similar assumption about the relationship between average expression and dropout rate [4]. Given a matrix, introduction of additional dropouts reduce average read count for each gene. On the flip side, using the above linear model, one can estimate D'_g associated with E'_g, where $E'_g = f \times E_g$. Here, $0 < f < 1$ is a factor that determines the decrease in average read-count, and is constant across all genes. Using Equation 4.9, one can compute the expected increase Δ_g in D_g due to change in E_g as follows.

$$\begin{aligned}
\Delta_g &= \widehat{D'_g} - \widehat{D_g} \\
&= [\alpha + \beta E'_g] - [\alpha + \beta E_g] \\
&= \beta(E'_g - E_g),
\end{aligned}$$

Here, $\widehat{D'_g}$ and $\widehat{D_g}$ are estimated log odds associated with E_g, and E'_g respectively. Also, $\Delta_g > 0$, since $\beta < 0$, and $E'_g < E_g$. We can utilise Δ_g to compute D'_g as follows.

$$D'_g = D_g + \Delta_g,$$

Finally, new dropout probability p'_g can be computed as follows.

$$p'_g = \frac{1}{1 + e^{-D'_g}}$$

We can also retrieve the updated average read count R'_g as using the below equation.

$$R'_g = 2^{E'_g} - 1,$$

Now, p'_g and R'_g are used to introduce additional dropouts, and adjusting average read count respectively. New drop outs are created by muting the expression of g in randomly chosen cells where it was earlier expressed. The number of additional dropouts can be easily calculated by tracking the difference between p'_g and p_g. After introducing the dropouts we calculate the interim average read count R^i_g. Further, we scale the cell-specific read counts of g by multiplying the values by $\frac{R'_g}{R^i_g}$.

4.1.7 Data and software availability

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO) under accession number GSE160910. The ROSeq R package is available at the Bioconductor portal. A more frequently updated version of the software can be accessed at: Github.

4.2 Results

4.2.1 Overview of ROSeq

Numerous parametric models have been proposed since the advent of single cell technology, primarily accounting for dropout cases. The majority of these are mixture models of distinct probability density functions. We conjecture that one of these approaches' limitations is that they disregard the noise and the technical bias, as commonly observed in single cell data.Ranks are commonly known to be more robust as compared to the corresponding expression estimates. In fact, with the increase in sample size, single cell studies are now seen embracing the traditional Wilcoxon's rank-sum test to identify differentially expressed genes. While non-parametric methods are assumption-free [189], they often lack statistical power. In this work, we explored the utility of discretizing an expression vector into bins and ordering them (meaning ordering ranks corresponding to bins) based on bin-wise cellular frequencies, thereby making it modellable by Discrete Generalized Beta Distribution (DGBD) (aka, rank-ordered distribution) [192].

Fitting DGBD on expression readouts involves MLE of two shape parameters, denoted by a and b. **(Figure 4.1B)** depicts an example of DGBD based modeling of (*VAMP3*) expression across 288 single cells from the biological replicate NA19098 of the Tung data [1] (For dataset description and naming convention refer to Materials and Methods). For a comprehensive assessment of the quality of fit, we estimated R^2 for all the 11513 genes that qualified the filtering criteria. DGBD fits yielded $R^2 > 0.9$ for a vast majority of the genes

(Figure 4.1C), thereby underscoring its appropriateness in modeling expression ranks. Leveraging DGBD based modeling of expression, we devised ROSeq, a Wald type test to determine differential expression in single cell data **(Figure 4.1A)**. We inspected the gene expression marginals (empirical distribution approximated using the *density* function by R) and the corresponding DGBD fits for some example DE/non-DE genes (called using ROSeq). We noticed that DGBD significantly stabilizes the shape diversity, as otherwise observed in case of the gene expression marginals. This highlights the strength of rank-ordered distribution, which homogenises diversely shaped marginals and enables reliable estimation of the distribution parameters.

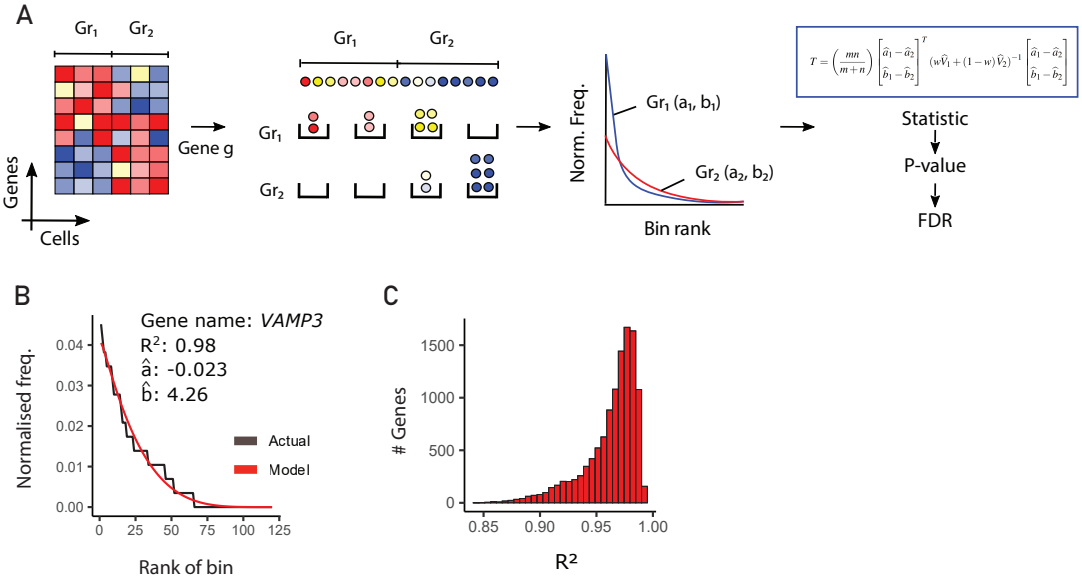

Figure 4.1: (A) As part of the ROSeq differential expression analysis workflow, cells are first binned depending on expression values associated with a particular gene. For each cell group, bins are ranked depending on cell frequency. The Discrete Generalized Beta Distribution (DGBD) is used as a probability mass function to express a normalized bin-wise cell-frequency as a function of its corresponding rank using two real parameters a and b. A Wald-type test is used on the MLE of these parameters across the cell-groups, to find differentially expressed genes. (B) Discrete Generalized Beta Distribution (DGBD) based modeling of *VAMP3* expression (Source: Tung data [1]). Discretized expression bins are ranked based on normalised bin-wise cellular frequencies. (C) Distribution of R^2 values obtained from DGBD based modeling of 11513 expressed genes (Source: Tung data [1]).

4.2.2 Comparative benchmarking based on matched bulk RNA sequencing data

When compared to single cell-based estimates, tissue-level gene expression measurements are considered to be more robust. As such, it's a common practice to benchmark single cell-based DEG calls against DEGs obtained from matched bulk expression profiles. We used scRNA-seq data from three previous studies that also performed bulk RNA-seq on the same samples. A total of eight contrasting cell-group pairs were constructed as follows — myoblasts before and 24 hours after differentiation (source: Trapnell data [193]), all three pairs of biological replicates of induced pluripotent stem cells (iPSCs) (source: Tung data [1]), and all three pairs of undifferentiated H1, H9 human embryonic stem cells (ESCs) and neuronal progenitor cells (NPCs) (source: Chu data [2]). We also profiled single expression of foreskin BJ fibroblasts and K562 cells, with matching bulk replicates (referred to as Gupta data). We used the standard Seurat pipeline (without batch correction) [203] to visualise the the single cells in the presence of batch information and obtained perfect segregation between the two cell types **(Figure 4.2)**, strengthening the case for straight-forward differential expression analysis.

Bulk replicates were used for confident DEG calls. In addition to ROSeq, single cell DEG calls were made using five best practice methods namely BPSC, SCDE, Wilcoxon's rank-sum test, MAST, and DESeq2 [196]. A single cell DEG call was considered true positive if the gene was also present in the list of DEGs detected by analysing the matching bulk-transcriptomes. If not, it was counted as a false positive. In six out of the eight cases, ROSeq topped

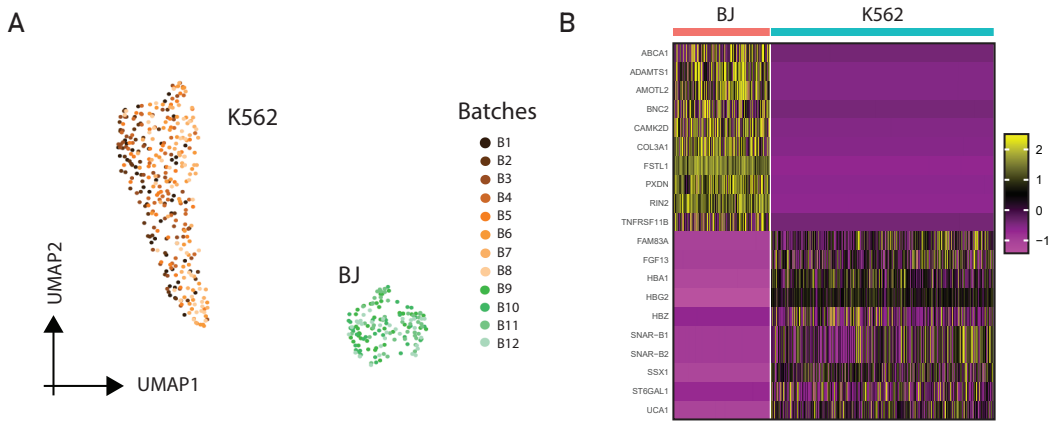

Figure 4.2: (A) Standard Seurat pipeline (without batch correction) was used for UMAP based visualisation of single cells, color-coded by batch information. Even in the absence of batch correction, cells look segregated by their respective types (brown dots represent K562 cells whereas green dots represent BJ fibroblast). (B) Heatmap showing top differentially expressed genes including known cell type markers such as FAM83A (found highly expressed in K562 cells), and COL3A1 (found highly expressed in BJ fibroblasts).

in terms of the estimated area under the ROC curve (AUC-ROC) values Figures 4.3 , 4.4a,B,E. SCDE performed best in the remaining two cases with a negligible margin over ROSeq **(Figure 4.4C,D)**. Although DESeq2 is not specialized for single cells, we used it as a control to ensure single cell focused methods yield overall better performance.

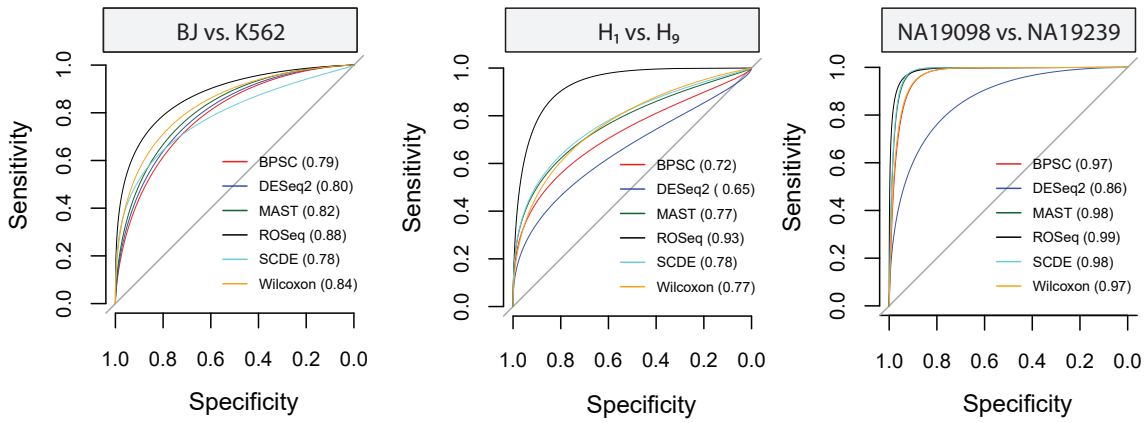

Figure 4.3: (A) ROC and the associated AUC values obtained by bulk-based benchmarking of single cell DEG calls between BJ and $K562$ cells (Gupta data). (B) ROC plot for H_1 and H_9 cells (Source: Chu Data [2]). (C) ROC plot for $NA19098$ and $NA19239$ cells (Source: Tung data [1]).

Besides AUC, we also tracked other popular measurements of classification

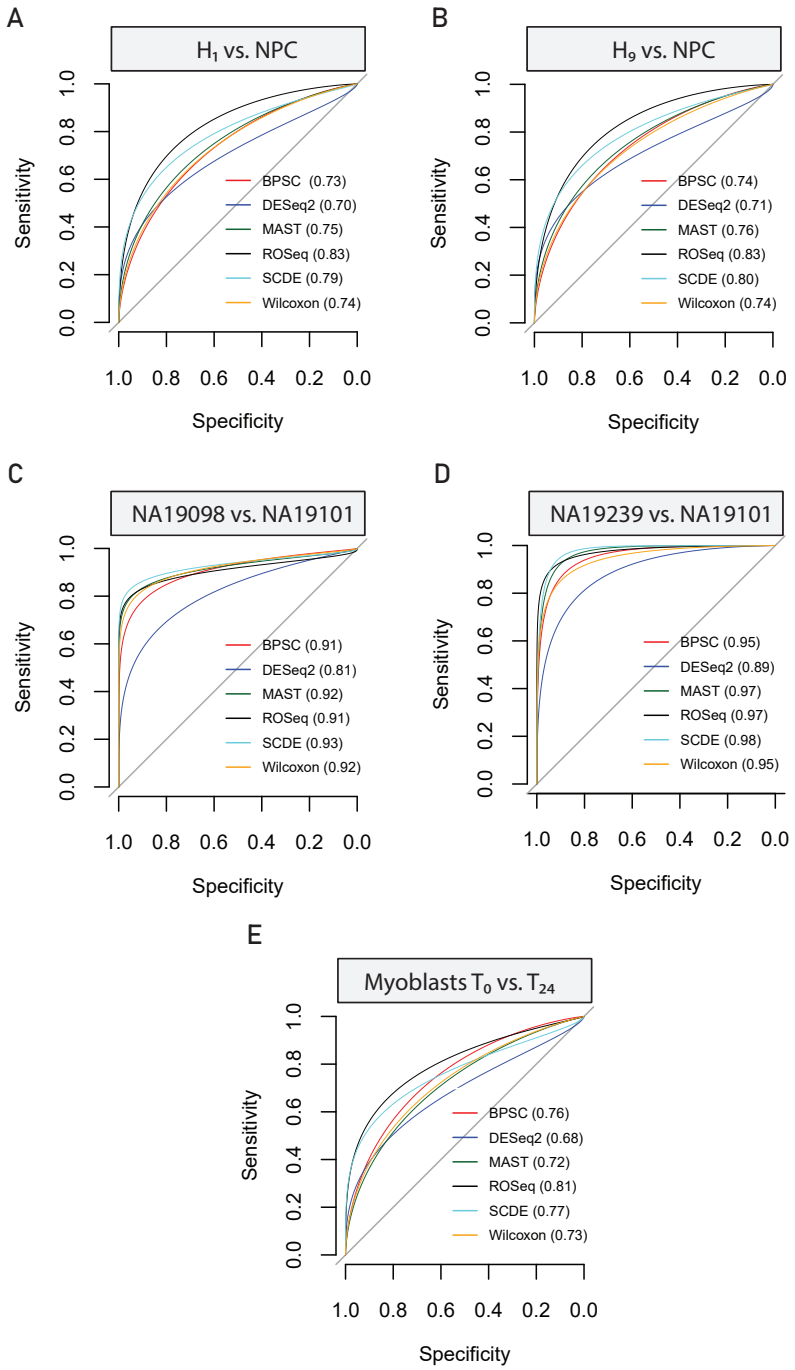

Figure 4.4: (A) Receiver Operating Characteristics curve (ROC) and the associated Area Under the Curve (AUC) values obtained by bulk-based benchmarking of single cell DEG calls between H1 and NPC cells (Chu data). (B) ROC plot for DEG calls between H9 and NPC (Chu data), (C) ROC plot for DEG calls between replicates NA19098 and NA19101 (Tung data). (D) ROC plot for DEG calls between replicates NA19239 and NA19101 (Tung data). (E) ROC plot for DEG calls between myoblasts sampled before and 24 hours after differentiation (Trapnell data).

accuracy, including F_1, Mathews Correlation Coefficient (MCC) [204], and Cohen's Kappa (k) [205]. Among these, MCC factors in the performance of a

binary classification system in all four confusion matrix categories, whereas Cohen's Kappa corrects the accuracy measurement by the expected performance. Of note, F_1, MCC and k were calculated on confusion matrices determined using a cut-off on the differential expression probabilities computed on the scRNA-seq datasets. Such a cut-off maximises the sum of sensitivity and specificity [206]. In the majority of the cases ROSeq maximised these scores, with striking margins in case of MCC and k. Based on the overall performance, the methods can be rank ordered as follows — ROSeq $\succ$ SCDE $\succ$ MAST $\succ$ Wilcoxon $\succ$ BPSC $\succ$ DESeq2.

ROSeq uses a constant k, which is multiplied with σ i.e. the standard deviation of the pooled expression estimates across the cell-groups (Materials and Methods). We observed that the choice of k impacts ROSeq's performance. On Gupta dataset comprising BJ fibroblasts and K562 cells, we assessed five different values of k — 0.01, 0.05, 0.1, 0.2, and 0.5. $k = 0.05$ stood out clearly, thereby strengthening its choice as a default **(Table 4.1)**.

Gupta; BJ, K562	**K=.01**	**K=.05**	**K=.1**	**K=.2**	**K=.5**
AUC	0.73	0.88	0.88	0.86	0.80
MCC	0.31	0.42	0.38	0.38	0.29
F1	0.35	0.43	0.38	0.39	0.31
Kappa	0.28	0.36	0.31	0.32	0.22

Table 4.1: ROSeq's performance (AUC, MCC, F1 and Kappa) on Gupta dataset, with values of k ranging from 0.01 to 0.5.

4.2.3 Type I errors

To evaluate ROSeq's Type I error control, we created several null datasets by dividing cells of the same type into two groups of different group sizes [195].

For each of the methods, we tracked the fraction of the tested genes that were assigned a nominal *P*-value. Three different cutoffs i.e., 0.01, 0.05, 0.1 were considered for the *P*-values. We iterated this simulation experiment for varied cell-group sizes — 50, 100, 200, 300, 400, 500, 1000, 1500. For each cell-group size, 20 null datasets were constructed and subjected to the DEG callers. For this experiment we used Jurkat transcriptomes from the Zheng data [3]. In addition to ROSeq, the performance of five other methods, namely BPSC, SCDE, Wilcoxon's rank-sum test, MAST, and DESeq2 were considered for comparison. Among all the six methods, ROSeq offered the overall best performance in all cases except for the cell groups having 100 or less number of cells **(Figure 4.5)**.

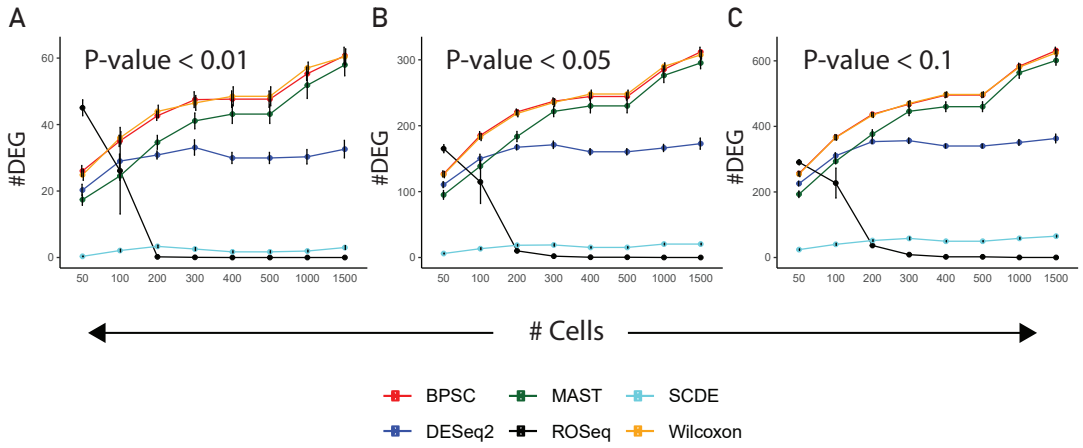

Figure 4.5: (A) Line chart showing Type I error rates with standard error (depicted by error bars), obtained by applying different DEG callers on 20 randomly sampled null datasets, for varied cell-group sizes. We applied a *P*-value cutoff of 0.01. These experiments were performed using Jurkat transcriptomes (3200 cells and 32000 transcripts [3]). (B, C) Similar plots with *P*-value cutoff of 0.05 and 0.1 repectively.

SCDE outperformed ROSeq with 50/100 cells in each group. The structure of rank-frequency distributions is better understood with more cells, as it aids in modelling the distribution spectrum in a finer grid. Furthermore, ROSeq's testing procedure uses asymptotic critical values (as determined by large-sample theory), resulting in improved inference for larger sample sizes. Prior to fit-

ting DGBD, ROSeq discretizes the observed expression by binning. As a result, as opposed to other methods, ROSeq needs to approximate gene expression marginals with smaller effective sample sizes. This, negatively influences the parameter estimation process, and can explain ROSeq's sub-optimal performance on smaller groups of cells. Rest of the methods including the successors of SCDE made significant number of DEG calls. ROSeq was found to be the only method which neared zero DEG calls with increased sample size. We tracked Standard Error (SE) scores across the various cases, which exhibited least variability to shuffling of the data.

4.2.4 Tolerance to noise due to excessive dropouts events

As previously mentioned, technological biases such as RNA degradation during cell isolation and processing, variable reagent numbers, the presence of cellular debris, and PCR amplification bias skew single cell gene expression readouts. Furthermore, even in the absence of technological variability, single cell expression estimates are inherently noisy due to the limited number of detected molecules. [189]. Majority of the state of the art dropout induction methods simulate scRNA-seq data with variable concentration of dropouts. This approach often involves making strong assumptions about gene expression marginals. We used the linear relationship between average read count and log odds of dropout rate, as discussed elsewhere, to establish a strategy for injecting dropouts into real scRNA-seq datasets [4]. This allowed us to introduce varied levels of dropouts by the means of controlling average read counts (Materials and Methods). We introduced various levels of dropouts (67-80%) in BJ fi-

broblasts and K562 cells. DEGs detected by analysing matching bulk RNA-seq dataset were used to compute AUC and MCC values. ROSeq clearly dominated the rest of the methods in calling the correct DEGs **(Figures 4.6A,B)**.

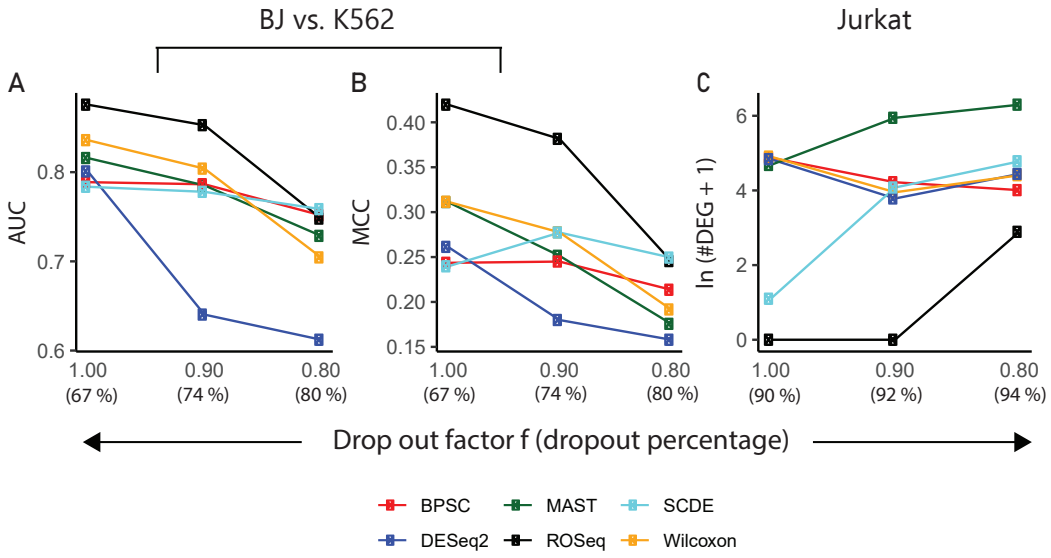

Figure 4.6: (A) Line chart showing decline in AUC with the increase in dropout levels. Performance was recorded on the Gupta dataset comprising BJ fibroblasts and K562 cells. (B) Line chart showing MCC values that largely mirror AUC values in subfigure A. (C) Line chart showing the trend of increased false DEG calls with the increase in dropout levels. Null datasets were created using Jurkat cell transcriptomes from the Zheng dataset. Each of the contrasting group contains 1000 cells.

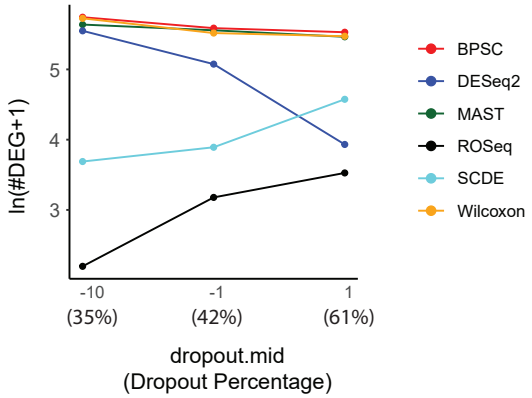

Figure 4.7: Line chart depicting false DEG calls on null dataset created using the splatter R package with different concentration of dropouts controlled by the custom parameter dropout.mid. Each simulated count matrix consisted of 6000 genes and 500 cells. Groups were created by randomly splitting the cells into two groups (250 cells in each group).

We also investigated the Type I errors by constructing null data by sampling Jurkat transcriptomes (source: Zheng dataset [3]). As we raised the dropout levels from 90 to 94 percent, ROSeq made the least amount of DEG calls, as

expected. **(Figure 4.6C)**. As an independent approach, we used the Splatter R package [4] to generate null datasets with variable dropout concentrations, which helped us to track the Type I error rates. ROSeq's performance remained consistent **(Figure 4.7)**. Collectively, these experiments reinforce the tolerance of ROSeq to noise caused by dropouts.

4.2.5 Runtime efficiency

The advent of droplet-based commercial platforms has enabled profiling of tens of thousands of cells in a single experiment has become a common affair. Unsupervised clustering of large scale scRNA-seq data produces numerous clusters, each of which typically harbors a large number of cells. As such, besides accuracy, the scalability has become a desirable feature for the DEG callers. We benchmarked time consumption by the methods for variable sizes of input scRNA-seq datasets. For the construction of the datasets, we performed the same steps as we did for estimating the Type I error rates. SCDE and BPSC are considerably slow as compared to the rest of the methods **(Figure 4.8A)**.

As a result,, we used a small dataset constituting 288 iPSCs (replicate id: NA19098) for tracking the execution time for all six methods (sampled scRNA-seq profiles with replacement due to lack of cells). This data consists of 19027 transcripts. ROSeq secured fourth place, following Wilcoxon, DESeq2 and MAST. SCDE was the slowest among all, followed by BPSC **(Figure 4.8A)**. To test on larger sample sizes, we made use of the Jurkat transcriptomes (source: Zheng data [3]) that allowed us to split the cells randomly into two equal sized groups (without replacement) with a maximum 1500 cells in each group. This

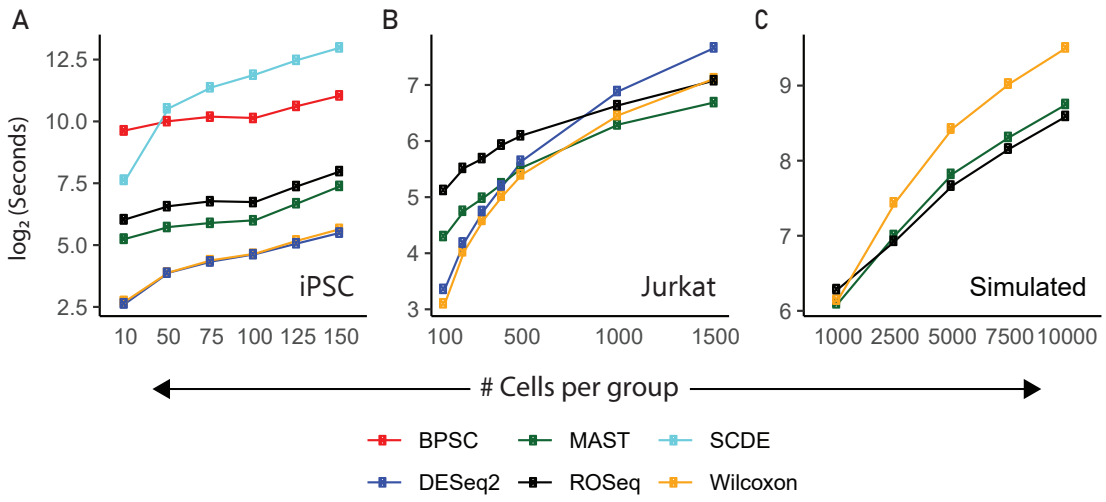

Figure 4.8: (A) Line chart showing median time taken by each algorithm on 100 randomly sampled null datasets containing iPSC transcriptomes (replicate id: NA19098). (B) Line chart showing median time taken by each algorithm on 20 randomly sampled null datasets containing Jurkat transcriptomes. (C) Line chart showing median time taken by each algorithm on 20 randomly sampled null datasets using the *Splatter* R package [4]. Note: For the iPSC data we used a single CPU core, for the remaining larger datsets we used 4 cores of the workstation.

data consists of 32738 transcripts. Owing to their longer processing times, SCDE and BPSC were dropped from the comparison. The number of cells in each group were varied from 100 to 1500. Although both of the approaches took almost the same amount of time, ROSeq took a downward turn as the number of cells increased. **(Figure 4.8B)**. This inspired us to speed-test the methods further on even larger samples sizes. To this end, we used the *Splatter* R package to simulate cells in up to 10000 sized groups (6000 genes). In this case, ROSeq turned out to be the fastest **(Figure 4.8C)**. MAST exhibited a similar performance, whereas Wilcoxon diverged significantly, thus suggesting some computing bottleneck. All experiments reported in this article were performed on a workstation configured with AMD Ryzen 7 3700X 8-Core processor with a clock speed of 4249.648 MHz, 64GB DDR4 RAM and Ubuntu 18.04.4 LTS operating system with 5.3.0-40-generic kernel. For the iPSC data, we used a single core, for the remaining larger datasets, we used 4 cores. We observed

that ROSeq speeds up significantly as number of cores are increased.

4.3 Discussion

Martinez-Mekler and colleagues revealed that a two-parameter DGBD (rank-ordered distribution) gives excellent fits to a wide range of phenomena in the arts, social sciences, and natural sciences. [192]. We evaluated the applicability of DGBD to gene expression data. We found DGBD to fit well to the entire spectrum of expressed genes of varying expression levels. We further developed ROSeq, a DGBD-based Wald type test for differential expression analysis of scRNA-seq data. Most of the statistical models for single cell expression data use mixed models to accommodate high dropout rates. ROSeq discretizes the data, thereby stabilizing local distortions in the shape of the distribution, due to noise and technical bias. Our experimentation with dropouts strengthens this conclusion. ROSeq exhibited best performance with the increase in artificially injected dropout levels. Most of the methods that rely on Negative Binomial or Poisson distributions enforce raw count data as input. ROSeq works on real values and does not impose such constraints. This is particularly beneficial since integrative single cell omics studies are very common these days, typically involving batch correction that inevitably transforms the read counts into real values. In this regard,it should be noted that ROSeq is not inbuilt with any batch correction method. As such, it expects the user to input an scRNA-seq dataset which is not only library size normalised but also free of other covariates as applicable.

We compared ROSeq's performance to some of the current best-practice methods, such as SCDE, MAST, and BPSC, which are primarily designed for single cell expression data. Among various critical observations, our systematic tracking of Type I errors showed that ROSeq, like SCDE, needs a comparatively larger number of cells (at least 100 in each contrasting group) to achieve optimal efficiency. Current studies record hundreds to thousands of cells per unsupervised cell cluster with the emergence of droplet-based single cell profiling platforms. Current studies report hundreds to thousands of cells per unsupervised cell cluster with the advent of droplet-based single cell profiling platforms. As such, we do not foresee any hindrance to ROSeq's applicability due to cell paucity. However, ROSeq might produce sub-optimal DEG calls if a cluster contains a small number of cells. Diverse types of progenitor cells, circulating tumor cells, etc., are examples of such rare cell types [112]. This shortcoming can be attributed to ROSeq's use of asymptotic distribution.

Comparing the marginal distribution of a gene's expression between two cell groups is a statistical test of differential expression. Across platforms, chemistry, and cellular conditions, gene expression marginals in single cells vary greatly. As such, it is difficult to rely on any specific parametric distribution function for modeling gene expression in single cells. Conversely, ROSeq analyzes the distribution of rank-ordered discretized expression bins across two cell-populations. We demonstrated that rank-ordered distribution stabilizes diversely shaped gene expression marginals Figures 4.9 , 4.10 while capturing necessary information about lineage/condition-specific expression patterns.

While ranks are considered to be lossy, they provide a means to bypass ex-

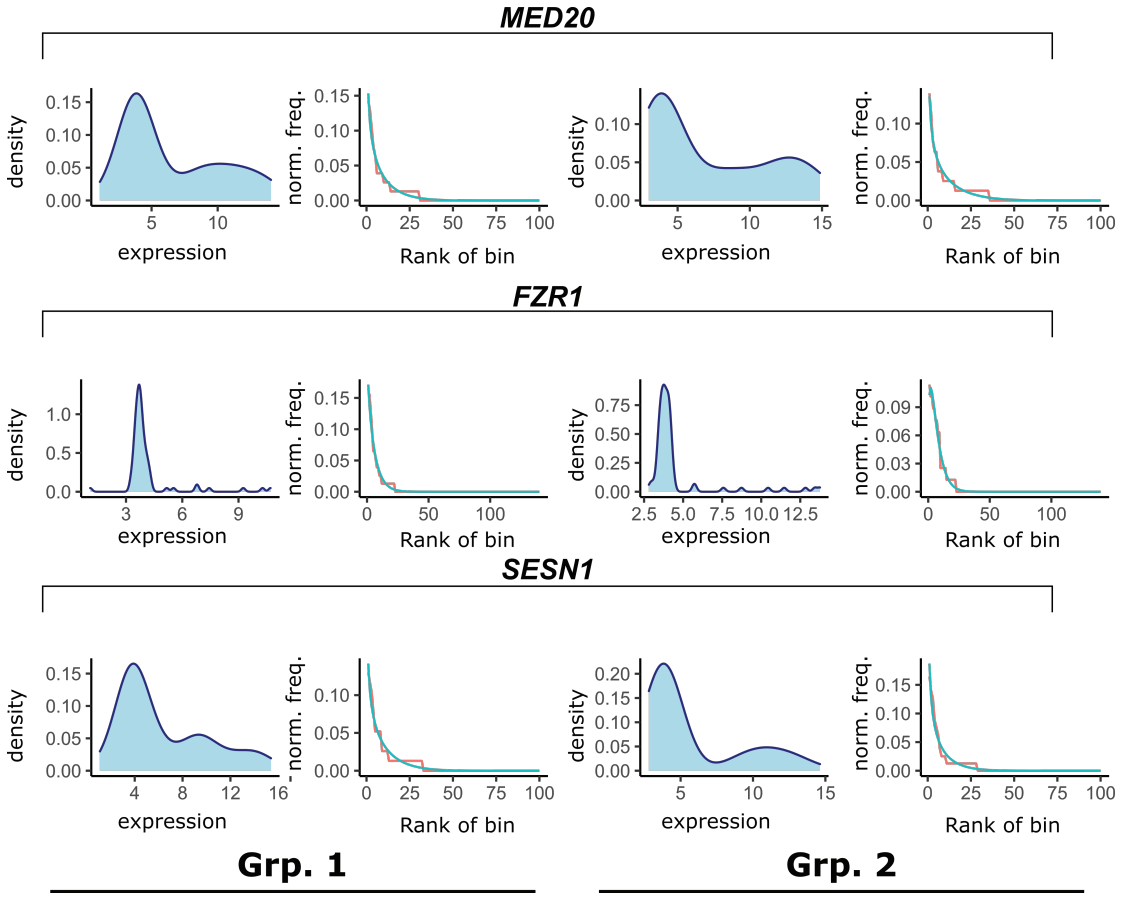

Figure 4.9: Five randomly selected non-differential genes between myoblasts sampled before/24 hours after differentiation (Source: Trapnell data). Differential expression analysis was performed using ROSeq. Each row contains cell-group-wise expression density plot as well as a plot depicting DGBD based modeling of rank-ordered expression bins.

pression modeling. The results presented in this work suggest that it could be beneficial to model gene expression ranks as compared to gene expression.

The approaches, which were optimised for single cell expression data, performed well when compared to bulk tissue-based DEG calls.However, SCDE and ROSeq maximized the DEG call accuracy. ROSeq is particularly powerful when the single cell based expression estimates are inherently noisy.In such instances, ROSeq significantly outperforms the other approaches that have been evaluated. Furthermore, since ROSeq seems to have a larger number of input

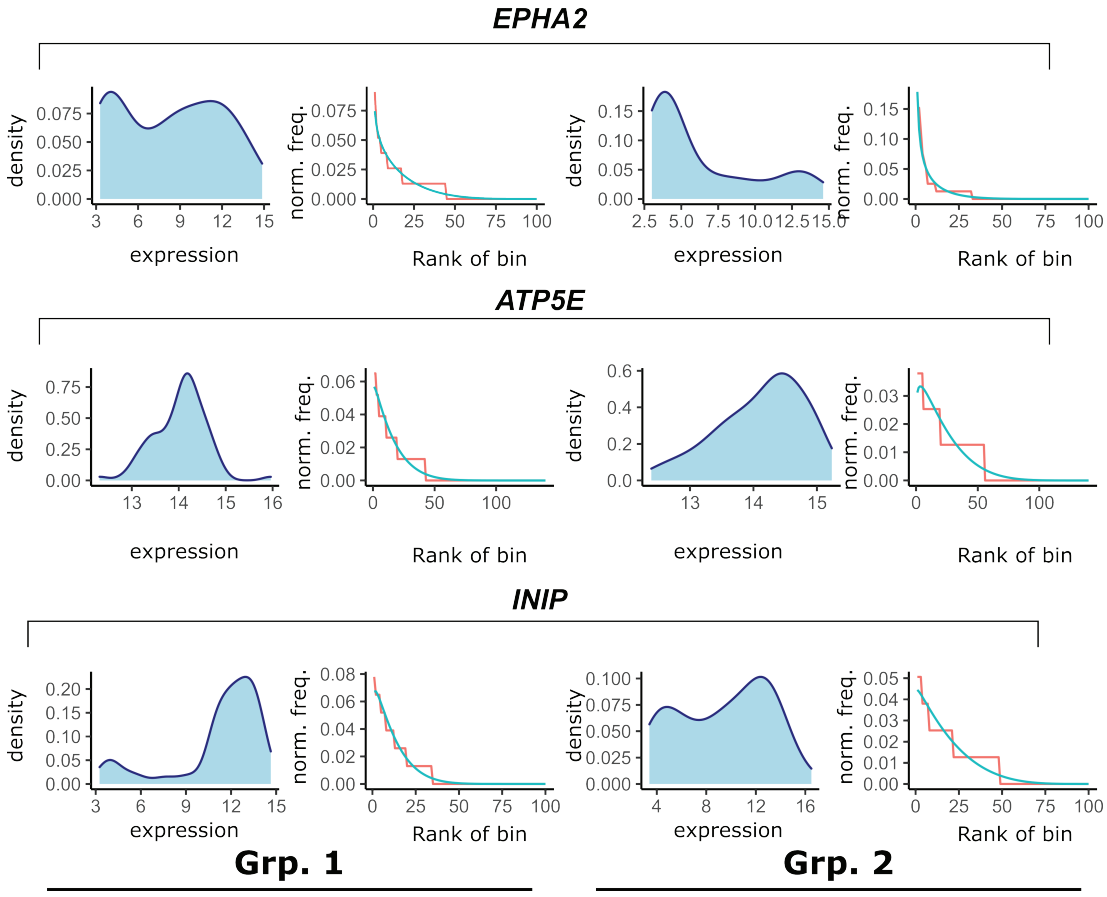

Figure 4.10: Five randomly selected differential genes between myoblasts sampled before/24 hours after differentiation (Source: Trapnell data). Differential expression analysis was performed using ROSeq. Each row contains cell-group-wise expression density plot as well as a plot depicting DGBD based modeling of rank-ordered expression bins.

cells, it has a faster turnaround time than other DEG callers. The current version of ROSeq allows for differential expression comparison between two classes of cells. With the aid of statistical inference theory, the Wald's statistic we used to formulate the differential expression test may be generalised to test the equality of parameter values across more than two independent classes.In this case, additional statistical tests would be required to pinpoint the parent cell-lineage associated with a DE gene. An immediate future extension of ROSeq, therefore would be enabling multi-group (≥ 2) comparisons.

Chapter 5

Conclusion

This thesis focuses on various mathematical modelling and bioinformatics analysis approaches of single-cell gene expression data. Our studies feature a significant element of integrative analysis of genomics data, curated from multiple studies. Further, we have demonstrated various applications of machine learning methods to predict disease states of cells.

5.1 Summary of contribution

In this section, we briefly summarise the chapter-wise contribution that gives the dissertation a bird's eye view.

5.1.1 Integrative Analysis and Machine Learning based Characterization of Single Circulating Tumor Cells (CTCs)

We incorporated single-cell expression profiles from various published studies and analyzed the emergence of epithelial to mesenchymal transition among single Circulating Tumor Cells (CTCs). We established the E: M score for this,

which allows the ordering of CTC transcriptomes on an estimated epithelial-mesenchymal transition pseudo-temporal axis. We leveraged the power of machine learning techniques in reliably distinguishing CTCs from other relatively way more abundant immune cell types. This is accomplished by incorporating CTC datasets that are publicly accessible and model training based on machine learning. We provide a user-friendly R package for CTC classification that provides a probabilistic score indicating the cancer origin of individual cells. Our reported Clear Cell® Polaris™ workflow, in tandem with the machine learning-based CTC-immune cell classification system, for the first time, enables truly unbiased detection of circulating tumor cells. We speculate a high adoption rate for our proposed strategy with declining per cell cost associated with single-cell gene expression screening.

5.1.2 The cellular basis of the loss of smell in 2019-nCoV infected individuals

Our study aims to underscore the potential cellular basis of the loss of olfaction, by examining the olfactory-epithelium-specific cell-types based on the expression levels of the host-specific viral-entry moieties as well as the burden of host-virus protein-protein interactions. Our findings indicate that the loss of odour in the infected patients may not be due to the clear damage of the sensory olfactory neurons. Rather olfactory stem cells (OSCs: HBCs and GBCs), sustentacular cells (SUSs), Bowman's gland cells, whose molecular make-up make the cells vulnerable to viral infection. This conclusion was drawn based on a consensus approach involving gene expression as well as host-pathogen protein interactome. Importantly, since all our findings are largely substantiated by in silico

analysis, one cannot obviate the limitations of the single-cell transcriptomics assay, such as sampling bias or high dropout rates [176]. Although the identification of susceptible cell types of the olfactory epithelium for the 2019-nCoV infection is defined based on the few identified viral entry moieties of the host cells, the potential involvement of currently uncharacterized host cell surface receptors or proteases that can promote viral entry into the host cells cannot be excluded.

5.1.3 ROSeq: Modelling expression ranks for noise-tolerant differential expression analysis of scRNA-Seq data

We tested DGBD's applicability to data on gene expression. We find that DGBD fits well with the entire spectrum of expressed genes with varying levels of expression. For differential expression analysis of scRNA-seq data, we further developed ROSeq, a DGBD-based Wald type test. Most of the single-cell expression data statistical models use mixed models to handle high dropout rates. ROSeq discretizes the data due to noise and technological bias, thus stabilizing local distortions in the form of the distribution. Our experimentation with dropouts confirms this conclusion. With the rise in the artificially injected dropout rate, ROSeq showed better results.

5.2 Future work

The computational methods presented in this thesis are not only limited to single-cell but can also be used for other data such as single-cell spatial tran-

scriptomics data. Spatial transcriptomics platforms feature ground breaking technology that provides both gene expression readouts and spatial location of cells. Certain variants of the spatial transcriptomics technologies provide 3D coordinates of each copy of a transcript inside a cell. The relative location of cells allows scientists to measure gene activity comprehensively within a tissue section. This allows critical investigation of disease pathogenesis in a spatial context. Precise tracking of transcript locations further allows scientists to find the heterogeneous orientations of mRNA accumulation and activities within seemingly similar cells. Below are some of the possible future extensions of the presented works.

- **ROseq for multiple groups :** The current version of ROSeq is limited to only two groups; It can be further extended to multiple groups. A suitable multi-group Wald test can be designed that performs differential expression analysis of multiple cellular clusters, as identified using single-cell clustering methods. We will also incorporate novel techniques to ward off batch effects.
- **Analysis of spatial transcriptomics data**
  - Distant cells have more probability of belonging to different cell types. Can we use this for type-2 error control in differential express genes analysis?
  - Can we estimate cellular organelles' morphological features that could help find heterogeneity within seemingly similar cells. We will make use of deep neural networks for this purpose.

- **Decline in transcriptional homeostasis defines aging** Age-dependent dysregulation of transcription regulatory machinery triggers modulations in the gene expression levels leading to the decline in cellular fitness. Tracking these transcripts along the temporal axis in multiple species revealed a spectrum of evolutionarily conserved pathways, such as electron transport chain, translation regulation, DNA repair, etc. Recent shreds of evidence suggest that aging deteriorates the transcription machinery itself, indicating the hidden complexity of the aging transcriptomes. This reinforces the need for devising novel computational methods to view aging through the lens of transcriptomics.

References

[1] P.-Y. Tung, J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K. Pritchard, and Y. Gilad, "Batch effects and the effective design of single-cell gene expression studies," *Scientific reports*, vol. 7, p. 39921, 2017.

[2] L.-F. Chu, N. Leng, J. Zhang, Z. Hou, D. Mamott, D. T. Vereide, J. Choi, C. Kendziorski, R. Stewart, and J. A. Thomson, "Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm," *Genome biology*, vol. 17, no. 1, p. 173, 2016.

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

[4] L. Zappia, B. Phipson, and A. Oshlack, "Splatter: simulation of single-cell rna sequencing data," *Genome biology*, vol. 18, no. 1, pp. 1–15, 2017.

[5] A. C. Pease, D. Solas, E. J. Sullivan, M. T. Cronin, C. P. Holmes, and S. Fodor, "Light-generated oligonucleotide arrays for rapid dna sequence

analysis," *Proceedings of the National Academy of Sciences*, vol. 91, no. 11, pp. 5022–5026, 1994.

[6] S. Nagpal, M. W. Karaman, M. M. Timmerman, V. V. Ho, B. L. Pike, and J. G. Hacia, "Improving the sensitivity and specificity of gene expression analysis in highly related organisms through the use of electronic masks," *Nucleic acids research*, vol. 32, no. 5, pp. e51–e51, 2004.

[7] C. Romualdi, S. Trevisan, B. Celegato, G. Costa, and G. Lanfranchi, "Improved detection of differentially expressed genes in microarray experiments through multiple scanning and image integration," *Nucleic Acids Research*, vol. 31, no. 23, pp. e149–e149, 2003.

[8] M. W. Pfaffl, "A new mathematical model for relative quantification in real-time rt–pcr," *Nucleic acids research*, vol. 29, no. 9, pp. e45–e45, 2001.

[9] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler, "Serial analysis of gene expression," *Science*, vol. 270, no. 5235, pp. 484–487, 1995.

[10] Y. Gitton, N. Dahmane, S. Baik, A. R. i Altaba, L. Neidhardt, M. Scholze, B. G. Herrmann, P. Kahlem, A. Benkahla, S. Schrinner *et al.*, "A gene expression map of human chromosome 21 orthologues in the mouse," *Nature*, vol. 420, no. 6915, pp. 586–591, 2002.

[11] L. Skrabanek and F. Campagne, "Tissueinfo: high-throughput identification of tissue expression profiles and specificity," *Nucleic Acids Research*, vol. 29, no. 21, pp. e102–e102, 2001.

[12] X. Mu, S. Zhao, R. Pershad, T.-F. Hsieh, A. Scarpa, S. W. Wang, R. A. White, P. D. Beremand, T. L. Thomas, L. Gan *et al.*, "Gene expression in the developing mouse retina by est sequencing and microarray analysis," *Nucleic acids research*, vol. 29, no. 24, pp. 4983–4993, 2001.

[13] R. Sorek and H. M. Safer, "A novel algorithm for computational identification of contaminated est libraries," *Nucleic Acids Research*, vol. 31, no. 3, pp. 1067–1074, 2003.

[14] T. Mahmood and P.-C. Yang, "Western blot: technique, theory, and trouble shooting," *North American journal of medical sciences*, vol. 4, no. 9, p. 429, 2012.

[15] G. S. Pall and A. J. Hamilton, "Improved northern blot method for enhanced detection of small rna," *Nature protocols*, vol. 3, no. 6, p. 1077, 2008.

[16] P. Argani, C. Iacobuzio-Donahue, B. Ryu, C. Rosty, M. Goggins, R. E. Wilentz, S. R. Murugesan, S. D. Leach, E. Jaffee, C. J. Yeo *et al.*, "Mesothelin is overexpressed in the vast majority of ductal adenocarcinomas of the pancreas: identification of a new pancreatic cancer marker by serial analysis of gene expression (sage)," *Clinical cancer research*, vol. 7, no. 12, pp. 3862–3868, 2001.

[17] F. Avila Cobos, J. Vandesompele, P. Mestdagh, and K. De Preter, "Computational deconvolution of transcriptomics data from mixed cell populations," *Bioinformatics*, vol. 34, no. 11, pp. 1969–1979, 2018.

[18] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[19] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome biology*, vol. 15, no. 12, p. 550, 2014.

[20] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for rna-sequencing and microarray studies," *Nucleic acids research*, vol. 43, no. 7, pp. e47–e47, 2015.

[21] S. S. Shen-Orr, R. Tibshirani, P. Khatri, D. L. Bodian, F. Staedtler, N. M. Perry, T. Hastie, M. M. Sarwal, M. M. Davis, and A. J. Butte, "Cell type–specific gene expression differences in complex tissues," *Nature methods*, vol. 7, no. 4, pp. 287–289, 2010.

[22] T. Gong and J. D. Szustakowski, "Deconrnaseq: a statistical framework for deconvolution of heterogeneous tissue samples based on mrna-seq data," *Bioinformatics*, vol. 29, no. 8, pp. 1083–1085, 2013.

[23] A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, and A. A. Alizadeh, "Robust enumeration of cell subsets from tissue expression profiles," *Nature methods*, vol. 12, no. 5, pp. 453–457, 2015.

[24] Y. Zhong, Y.-W. Wan, K. Pang, L. M. Chow, and Z. Liu, "Digital sorting of complex tissues for cell type-specific gene expression profiles," *BMC bioinformatics*, vol. 14, no. 1, p. 89, 2013.

[25] K. J. Livak and T. D. Schmittgen, "Analysis of relative gene expression data using real-time quantitative pcr and the 2- δ δ ct method," *methods*, vol. 25, no. 4, pp. 402–408, 2001.

[26] C. A. Maher, C. Kumar-Sinha, X. Cao, S. Kalyana-Sundaram, B. Han, X. Jing, L. Sam, T. Barrette, N. Palanisamy, and A. M. Chinnaiyan, "Transcriptome sequencing to detect gene fusions in cancer," *Nature*, vol. 458, no. 7234, pp. 97–101, 2009.

[27] N. T. Ingolia, G. A. Brar, S. Rouskin, A. M. McGeachy, and J. S. Weissman, "The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mrna fragments," *Nature protocols*, vol. 7, no. 8, pp. 1534–1550, 2012.

[28] K. R. Kukurba and S. B. Montgomery, "Rna sequencing and analysis," *Cold Spring Harbor Protocols*, vol. 2015, no. 11, pp. pdb–top084 970, 2015.

[29] P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss *et al.*, "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics," *Science*, vol. 353, no. 6294, pp. 78–82, 2016.

[30] R. Bacher and C. Kendziorski, "Design and computational analysis of single-cell rna-sequencing experiments," *Genome biology*, vol. 17, no. 1, p. 63, 2016.

[31] G. K. Marinov, B. A. Williams, K. McCue, G. P. Schroth, J. Gertz, R. M. Myers, and B. J. Wold, "From single-cell to cell-pool transcriptomes: stochasticity in gene expression and rna splicing," *Genome research*, vol. 24, no. 3, pp. 496–510, 2014.

[32] D. Grün, L. Kester, and A. Van Oudenaarden, "Validation of noise models for single-cell transcriptomics," *Nature methods*, vol. 11, no. 6, p. 637, 2014.

[33] P. Brennecke, S. Anders, J. K. Kim, A. A. Kołodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni *et al.*, "Accounting for technical noise in single-cell rna-seq experiments," *Nature methods*, vol. 10, no. 11, pp. 1093–1095, 2013.

[34] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, "Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells," *Nature biotechnology*, vol. 33, no. 2, pp. 155–160, 2015.

[35] C. A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, and J. C. Marioni, "Normalizing single-cell rna sequencing data: challenges and opportunities," *Nature methods*, vol. 14, no. 6, p. 565, 2017.

[36] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, "Bayesian approach to single-cell differential expression analysis," *Nature methods*, vol. 11, no. 7, p. 740, 2014.

[37] 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.

[38] S. C. Hicks, F. W. Townes, M. Teng, and R. A. Irizarry, "Missing data and technical variability in single-cell rna-sequencing experiments," *Biostatistics*, vol. 19, no. 4, pp. 562–578, 2018.

[39] A. Tanay and A. Regev, "Scaling single-cell genomics from phenomenology to mechanism," *Nature*, vol. 541, no. 7637, pp. 331–338, 2017.

[40] X. Han, Z. Zhou, L. Fei, H. Sun, R. Wang, Y. Chen, H. Chen, J. Wang, H. Tang, W. Ge *et al.*, "Construction of a human cell landscape at single-cell level," *Nature*, vol. 581, no. 7808, pp. 303–309, 2020.

[41] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys, "A comparison of single-cell trajectory inference methods," *Nature biotechnology*, vol. 37, no. 5, pp. 547–554, 2019.

[42] X.-X. Chen and F. Bai, "Single-cell analyses of circulating tumor cells," *Cancer biology & medicine*, vol. 12, no. 3, p. 184, 2015.

[43] A. F. Sarioglu, N. Aceto, N. Kojic, M. C. Donaldson, M. Zeinali, B. Hamza, A. Engstrom, H. Zhu, T. K. Sundaresan, D. T. Miyamoto

et al., "A microfluidic device for label-free, physical capture of circulating tumor cell clusters," *Nature methods*, vol. 12, no. 7, p. 685, 2015.

[44] M. E. Warkiani, G. Guan, K. B. Luan, W. C. Lee, A. A. S. Bhagat, P. K. Chaudhuri, D. S.-W. Tan, W. T. Lim, S. C. Lee, P. C. Chen *et al.*, "Slanted spiral microfluidics for the ultra-fast, label-free isolation of circulating tumor cells," *Lab on a Chip*, vol. 14, no. 1, pp. 128–137, 2014.

[45] N. M. Karabacak, P. S. Spuhler, F. Fachin, E. J. Lim, V. Pai, E. Ozkumur, J. M. Martel, N. Kojic, K. Smith, P.-i. Chen *et al.*, "Microfluidic, marker-free isolation of circulating tumor cells from blood samples," *Nature protocols*, vol. 9, no. 3, p. 694, 2014.

[46] L. Xu, X. Mao, A. Imrali, F. Syed, K. Mutsvangwa, D. Berney, P. Cathcart, J. Hines, J. Shamash, and Y.-J. Lu, "Optimization and evaluation of a novel size based circulating tumor cell isolation system," *PloS one*, vol. 10, no. 9, p. e0138032, 2015.

[47] M. E. Warkiani, B. L. Khoo, L. Wu, A. K. P. Tay, A. A. S. Bhagat, J. Han, and C. T. Lim, "Ultra-fast, label-free isolation of circulating tumor cells from blood using spiral microfluidics," *Nature protocols*, vol. 11, no. 1, p. 134, 2016.

[48] C. Dive and G. Brady, "Snapshot: circulating tumor cells," *Cell*, vol. 168, no. 4, pp. 742–742, 2017.

[49] A. Marusyk and K. Polyak, "Tumor heterogeneity: causes and consequences," *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, vol. 1805, no. 1, pp. 105–117, 2010.

[50] J. G. Gall, "Human genome sequencing," *Science*, vol. 233, no. 4771, pp. 1367–1368, 1986.

[51] J. G. Reiter, A. P. Makohon-Moore, J. M. Gerold, A. Heyde, M. A. Attiyeh, Z. A. Kohutek, C. J. Tokheim, A. Brown, R. M. DeBlasio, J. Niyazov *et al.*, "Minimal functional driver gene heterogeneity among untreated metastases," *Science*, vol. 361, no. 6406, pp. 1033–1037, 2018.

[52] G. H. Heppner, "Tumor heterogeneity," *Cancer research*, vol. 44, no. 6, pp. 2259–2265, 1984.

[53] D. R. Welch, "Tumor heterogeneity—a 'contemporary concept'founded on historical insights and predictions," *Cancer research*, vol. 76, no. 1, pp. 4–6, 2016.

[54] L. S. Lindstrom, E. Karlsson, U. M. Wilking, U. Johansson, J. Hartman, E. K. Lidbrink, T. Hatschek, L. Skoog, and J. Bergh, "Clinically used breast cancer markers such as estrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2 are unstable throughout tumor progression," *J Clin Oncol*, vol. 30, no. 21, pp. 2601–2608, 2012.

[55] M. Gerlinger, A. J. Rowan, S. Horswell, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey *et al.*, "Intratumor heterogeneity and branched evolution revealed by multiregion sequencing," *N Engl j Med*, vol. 366, pp. 883–892, 2012.

[56] J. Massagué and A. C. Obenauf, "Metastatic colonization by circulating tumour cells," *Nature*, vol. 529, no. 7586, pp. 298–306, 2016.

[57] X. Liu, R. Taftaf, M. Kawaguchi, Y.-F. Chang, W. Chen, D. Entenberg, Y. Zhang, L. Gerratana, S. Huang, D. B. Patel *et al.*, "Homophilic cd44 interactions mediate tumor cell aggregation and polyclonal metastasis in patient-derived breast cancer models," *Cancer discovery*, vol. 9, no. 1, pp. 96–113, 2019.

[58] D. A. Lawson, K. Kessenbrock, R. T. Davis, N. Pervolarakis, and Z. Werb, "Tumour heterogeneity and metastasis at single-cell resolution," *Nature cell biology*, vol. 20, no. 12, pp. 1349–1360, 2018.

[59] M. Riquet, C. Rivera, L. Gibault, C. Pricopi, P. Mordant, A. Badia, A. Arame, and F. P. B. Le, "Lymphatic spread of lung cancer: anatomical lymph node chains unchained in zones," *Revue de pneumologie clinique*, vol. 70, no. 1-2, pp. 16–25, 2014.

[60] L. M. Millner, M. W. Linder, and R. Valdes, "Circulating tumor cells: a review of present methods and the need to identify heterogeneous phenotypes," *Annals of Clinical & Laboratory Science*, vol. 43, no. 3, pp. 295–304, 2013.

[61] F. Fabbri, S. Carloni, W. Zoli, P. Ulivi, G. Gallerani, P. Fici, E. Chiadini, A. Passardi, G. L. Frassineti, A. Ragazzini *et al.*, "Detection and recovery of circulating colon cancer cells using a dielectrophoresis-based device: Kras mutation status in pure ctcs," *Cancer letters*, vol. 335, no. 1, pp. 225–231, 2013.

[62] A. Ståhlberg and M. Kubista, "The workflow of single-cell expression profiling using quantitative real-time pcr," *Expert review of molecular*

diagnostics, vol. 14, no. 3, pp. 323–331, 2014.

[63] A. Toss, Z. Mu, S. Fernandez, and M. Cristofanilli, "Ctc enumeration and characterization: moving toward personalized medicine," *Annals of translational medicine*, vol. 2, no. 11, 2014.

[64] A. Chess, "Monoallelic gene expression in mammals," *Annual Review of Genetics*, vol. 50, pp. 317–327, 2016.

[65] A. S. Garfield, M. Cowley, F. M. Smith, K. Moorwood, J. E. Stewart-Cox, K. Gilroy, S. Baker, J. Xia, J. W. Dalley, L. D. Hurst *et al.*, "Distinct physiological and behavioural functions for parental alleles of imprinted grb10," *Nature*, vol. 469, no. 7331, pp. 534–538, 2011.

[66] B. H. Mecham, G. T. Klus, J. Strovel, M. Augustus, D. Byrne, P. Bozso, D. Z. Wetmore, T. J. Mariani, I. S. Kohane, and Z. Szallasi, "Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements," *Nucleic Acids Research*, vol. 32, no. 9, pp. e74–e74, 2004.

[67] M. Kaern, T. C. Elston, W. J. Blake, and J. J. Collins, "Stochasticity in gene expression: from theories to phenotypes," *Nature Reviews Genetics*, vol. 6, no. 6, pp. 451–464, 2005.

[68] N. Maheshri and E. K. O'Shea, "Living with noisy genes: how cells function reliably with inherent variability in gene expression," *Annual review of biophysics and biomolecular structure*, vol. 36, 2007.

[69] N. Leng, L.-F. Chu, C. Barry, Y. Li, J. Choi, X. Li, P. Jiang, R. M. Stewart, J. A. Thomson, and C. Kendziorski, "Oscope identifies oscillatory

genes in unsynchronized single-cell rna-seq experiments," *Nature methods*, vol. 12, no. 10, p. 947, 2015.

[70] T. M. A. Neildez-Nguyen, A. Parisot, C. Vignal, P. Rameau, D. Stockholm, J. Picot, V. Allo, C. Le Bec, C. Laplace, and A. Paldi, "Epigenetic gene expression noise and phenotypic diversification of clonal cell populations," *Differentiation*, vol. 76, no. 1, pp. 33–40, 2008.

[71] D. Sinha, A. Kumar, H. Kumar, S. Bandyopadhyay, and D. Sengupta, "dropclust: efficient clustering of ultra-large scrna-seq data," *Nucleic acids research*, vol. 46, no. 6, pp. e36–e36, 2018.

[72] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck *et al.*, "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets," *Cell*, vol. 161, no. 5, pp. 1202–1214, 2015.

[73] A. Zeisel, W. J. Köstler, N. Molotski, J. M. Tsai, R. Krauthgamer, J. Jacob-Hirsch, G. Rechavi, Y. Soen, S. Jung, Y. Yarden *et al.*, "Coupled pre-mrna and mrna dynamics unveil operational strategies underlying transcriptional responses to stimuli," *Molecular systems biology*, vol. 7, no. 1, p. 529, 2011.

[74] J. M. Gray, D. A. Harmin, S. A. Boswell, N. Cloonan, T. E. Mullen, J. J. Ling, N. Miller, S. Kuersten, Y.-C. Ma, S. A. McCarroll *et al.*, "Snapshot-seq: a method for extracting genome-wide, in vivo mrna dynamics from a single total rna sample," *PloS one*, vol. 9, no. 2, p. e89673, 2014.

[75] S. Linnarsson and S. A. Teichmann, "Single-cell genomics: coming of age," 2016.

[76] G. M. d. Oliveira, M. I. D. Rossi *et al.*, "The possible management of hypertensive patients infected with covid-19 through the use of aliskiren," 2020.

[77] T. Mashimo, K. Pichumani, V. Vemireddy, K. J. Hatanpaa, D. K. Singh, S. Sirasanagandla, S. Nannepaga, S. G. Piccirillo, Z. Kovacs, C. Foong *et al.*, "Acetate is a bioenergetic substrate for human glioblastoma and brain metastases," *Cell*, vol. 159, no. 7, pp. 1603–1614, 2014.

[78] N. Rahman, Z. Basharat, M. Yousuf, G. Castaldo, L. Rastrelli, and H. Khan, "Virtual screening of natural products against type ii transmembrane serine protease (tmprss2), the priming agent of coronavirus 2 (sars-cov-2)," *Molecules*, vol. 25, no. 10, p. 2271, 2020.

[79] A. G. Tibbe, M. C. Miller, and L. W. Terstappen, "Statistical considerations for enumeration of circulating tumor cells," *Cytometry Part A: the journal of the International Society for Analytical Cytology*, vol. 71, no. 3, pp. 154–162, 2007.

[80] D. Di Carlo, "Inertial microfluidics," *Lab on a Chip*, vol. 9, no. 21, pp. 3038–3046, 2009.

[81] K. Loutherback, J. D'Silva, L. Liu, A. Wu, R. H. Austin, and J. C. Sturm, "Deterministic separation of cancer cells from blood at 10 ml/min," *AIP advances*, vol. 2, no. 4, p. 042107, 2012.

[82] G. Barriere, P. Fici, G. Gallerani, F. Fabbri, W. Zoli, and M. Rigaud, "Circulating tumor cells and epithelial, mesenchymal and stemness markers: characterization of cell subpopulations," *Annals of translational medicine*, vol. 2, no. 11, 2014.

[83] T. M. Gall and A. E. Frampton, "Gene of the month: E-cadherin (cdh1)," *Journal of clinical pathology*, vol. 66, no. 11, pp. 928–932, 2013.

[84] S. V. Litvinov, M. P. Velders, H. Bakker, G. J. Fleuren, and S. O. Warnaar, "Ep-cam: a human epithelial antigen is a homophilic cell-cell adhesion molecule." *The Journal of cell biology*, vol. 125, no. 2, pp. 437–446, 1994.

[85] J. P. Thiery and J. P. Sleeman, "Complex networks orchestrate epithelial–mesenchymal transitions," *Nature reviews Molecular cell biology*, vol. 7, no. 2, pp. 131–142, 2006.

[86] E. A. Runkle and D. Mu, "Tight junction proteins: from barrier to tumorigenesis," *Cancer letters*, vol. 337, no. 1, pp. 41–48, 2013.

[87] R. B. Hazan, R. Qiao, R. Keren, I. Badano, and K. Suyama, "Cadherin switch in tumor progression," *Annals of the New York Academy of Sciences*, vol. 1014, no. 1, pp. 155–163, 2004.

[88] H. J. Hugo, L. Pereira, R. Suryadinata, Y. Drabsch, T. J. Gonda, N. D. Gunasinghe, C. Pinto, E. T. Soo, B. J. van Denderen, P. Hill *et al.*, "Direct repression of myb by zeb1 suppresses proliferation and epithelial gene expression during epithelial-to-mesenchymal transition of breast cancer cells," *Breast cancer research*, vol. 15, no. 6, p. R113, 2013.

[89] C. C. Warzecha and R. P. Carstens, "Complex changes in alternative pre-mrna splicing play a central role in the epithelial-to-mesenchymal transition (emt)," in *Seminars in cancer biology*, vol. 22, no. 5-6. Elsevier, 2012, pp. 417–427.

[90] M. S. Pino, M. Balsamo, F. Di Modugno, M. Mottolese, M. Alessio, E. Melucci, M. Milella, D. J. McConkey, U. Philippar, F. B. Gertler *et al.*, "Human mena+ 11a isoform serves as a marker of epithelial phenotype and sensitivity to epidermal growth factor receptor inhibition in human pancreatic cancer cell lines," *Clinical Cancer Research*, vol. 14, no. 15, pp. 4943–4950, 2008.

[91] M. Balic, H. Lin, L. Young, D. Hawes, A. Giuliano, G. McNamara, R. H. Datar, and R. J. Cote, "Most early disseminated cancer cells detected in bone marrow of breast cancer patients have a putative breast cancer stem cell phenotype," *Clinical cancer research*, vol. 12, no. 19, pp. 5615–5621, 2006.

[92] T. N. Seyfried and L. C. Huysentruyt, "On the origin of cancer metastasis," *Critical reviews in oncogenesis*, vol. 18, no. 1-2, p. 43, 2013.

[93] Y. Song, T. Tian, Y. Shi, W. Liu, Y. Zou, T. Khajvand, S. Wang, Z. Zhu, and C. Yang, "Enrichment and single-cell analysis of circulating tumor cells," *Chemical science*, vol. 8, no. 3, pp. 1736–1751, 2017.

[94] E. Andreopoulou, L.-Y. Yang, K. Rangel, J. Reuben, L. Hsu, S. Krishnamurthy, V. Valero, H. Fritsche, and M. Cristofanilli, "Comparison of assay methods for detection of circulating tumor cells in metastatic breast

cancer: Adnagen adnatest breastcancer select/detect™ versus veridex cellsearch™ system," *International journal of cancer*, vol. 130, no. 7, pp. 1590–1597, 2012.

[95] S. D. Mikolajczyk, L. S. Millar, P. Tsinberg, S. M. Coutts, M. Zomorrodi, T. Pham, F. Z. Bischoff, and T. J. Pircher, "Detection of epcam-negative and cytokeratin-negative circulating tumor cells in peripheral blood," *Journal of oncology*, vol. 2011, 2011.

[96] M. C. Miller, G. V. Doyle, and L. W. Terstappen, "Significance of circulating tumor cells detected by the cellsearch system in patients with metastatic breast colorectal and prostate cancer," *Journal of oncology*, vol. 2010, 2010.

[97] F. Farace, C. Massard, N. Vimond, F. Drusch, N. Jacques, F. Billiot, A. Laplanche, A. Chauchereau, L. Lacroix, D. Planchard *et al.*, "A direct comparison of cellsearch and iset for circulating tumour-cell detection in patients with metastatic carcinomas," *British journal of cancer*, vol. 105, no. 6, pp. 847–853, 2011.

[98] L. Wang, P. Balasubramanian, A. P. Chen, S. Kummar, Y. A. Evrard, and R. J. Kinders, "Promise and limits of the cellsearch platform for evaluating pharmacodynamics in circulating tumor cells," in *Seminars in oncology*, vol. 43. Elsevier, 2016, pp. 464–475.

[99] M. T. Gabriel, L. R. Calleja, A. Chalopin, B. Ory, and D. Heymann, "Circulating tumor cells: a review of non–epcam-based approaches for cell

enrichment and isolation," *Clinical chemistry*, vol. 62, no. 4, pp. 571–581, 2016.

[100] M. M. Ferreira, V. C. Ramani, and S. S. Jeffrey, "Circulating tumor cell technologies," *Molecular oncology*, vol. 10, no. 3, pp. 374–394, 2016.

[101] Y.-H. Cheng, Y.-C. Chen, E. Lin, R. Brien, S. Jung, Y.-T. Chen, W. Lee, Z. Hao, S. Sahoo, H. M. Kang *et al.*, "Hydro-seq enables contamination-free high-throughput single-cell rna-sequencing for circulating tumor cells," *Nature Communications*, vol. 10, no. 1, p. 2163, 2019.

[102] N. Aceto, A. Bardia, D. T. Miyamoto, M. C. Donaldson, B. S. Wittner, J. A. Spencer, M. Yu, A. Pely, A. Engstrom, H. Zhu *et al.*, "Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis," *Cell*, vol. 158, no. 5, pp. 1110–1122, 2014.

[103] E. Lin, T. Cao, S. Nagrath, and M. R. King, "Circulating tumor cells: diagnostic and therapeutic applications," *Annual review of biomedical engineering*, vol. 20, pp. 329–352, 2018.

[104] N. Aceto, A. Bardia, B. S. Wittner, M. C. Donaldson, R. O'Keefe, A. Engstrom, F. Bersani, Y. Zheng, V. Comaills, K. Niederhoffer *et al.*, "Ar expression in breast cancer ctcs associates with bone metastases," *Molecular Cancer Research*, vol. 16, no. 4, pp. 720–727, 2018.

[105] Y. Zheng, D. T. Miyamoto, B. S. Wittner, J. P. Sullivan, N. Aceto, N. V. Jordan, M. Yu, N. M. Karabacak, V. Comaills, R. Morris *et al.*, "Expression of β-globin by cancer cells promotes cell survival during blood-borne dissemination," *Nature communications*, vol. 8, p. 14344, 2017.

[106] D. T. Ting, B. S. Wittner, M. Ligorio, N. V. Jordan, A. M. Shah, D. T. Miyamoto, N. Aceto, F. Bersani, B. W. Brannigan, K. Xega *et al.*, "Single-cell rna sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells," *Cell reports*, vol. 8, no. 6, pp. 1905–1918, 2014.

[107] D. T. Miyamoto, Y. Zheng, B. S. Wittner, R. J. Lee, H. Zhu, K. T. Broderick, R. Desai, D. B. Fox, B. W. Brannigan, J. Trautwein *et al.*, "Rna-seq of single prostate ctcs implicates noncanonical wnt signaling in antiandrogen resistance," *Science*, vol. 349, no. 6254, pp. 1351–1356, 2015.

[108] M. G. van der Wijst, H. Brugge, D. H. de Vries, P. Deelen, M. A. Swertz, and L. Franke, "Single-cell rna sequencing identifies celltype-specific cis-eqtls and co-expression qtls," *Nature genetics*, vol. 50, no. 4, p. 493, 2018.

[109] N. V. Jordan, A. Bardia, B. S. Wittner, C. Benes, M. Ligorio, Y. Zheng, M. Yu, T. K. Sundaresan, J. A. Licausi, R. Desai *et al.*, "Her2 expression identifies dynamic functional states within circulating breast cancer cells," *Nature*, vol. 537, no. 7618, p. 102, 2016.

[110] S. Gkountela, F. Castro-Giner, B. M. Szczerba, M. Vetter, J. Landin, R. Scherrer, I. Krol, M. C. Scheidmann, C. Beisel, C. U. Stirnimann *et al.*, "Circulating tumor cell clustering shapes dna methylation to enable metastasis seeding," *Cell*, vol. 176, no. 1-2, pp. 98–112, 2019.

[111] B. M. Szczerba, F. Castro-Giner, M. Vetter, I. Krol, S. Gkountela, J. Landin, M. C. Scheidmann, C. Donato, R. Scherrer, J. Singer *et al.*,

"Neutrophils escort circulating tumour cells to enable cell cycle progression," *Nature*, p. 1, 2019.

[112] A. Jindal, P. Gupta, D. Sengupta *et al.*, "Discovery of rare cells from voluminous single cell expression data," *Nature communications*, vol. 9, no. 1, pp. 1–9, 2018.

[113] D. Srivastava, A. Iyer, V. Kumar, and D. Sengupta, "Cellatlassearch: a scalable search engine for single cells," *Nucleic acids research*, vol. 46, no. W1, pp. W141–W147, 2018.

[114] D. Sinha, P. Sinha, R. Saha, S. Bandyopadhyay, and D. Sengupta, "Improved dropclust r package with integrative analysis support for scrna-seq data," 2020.

[115] X. Zhang, Y. Lan, J. Xu, F. Quan, E. Zhao, C. Deng, T. Luo, L. Xu, G. Liao, M. Yan *et al.*, "Cellmarker: a manually curated resource of cell markers in human and mouse," *Nucleic acids research*, vol. 47, no. D1, pp. D721–D728, 2018.

[116] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[117] I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-r. Loh, and S. Raychaudhuri, "Fast, sensitive and accurate integration of single-cell data with harmony," *Nature methods*, pp. 1–8, 2019.

[118] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[119] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[120] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[121] H. Li, E. T. Courtois, D. Sengupta, Y. Tan, K. H. Chen, J. J. L. Goh, S. L. Kong, C. Chua, L. K. Hon, W. S. Tan *et al.*, "Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors," *Nature genetics*, vol. 49, no. 5, p. 708, 2017.

[122] M. A. Nieto, R. Y.-J. Huang, R. A. Jackson, and J. P. Thiery, "Emt: 2016," *Cell*, vol. 166, no. 1, pp. 21–45, 2016.

[123] C. M. Bailey, J. A. Morrison, and P. M. Kulesa, "Melanoma revives an embryonic migration program to promote plasticity and invasion," *Pigment cell & melanoma research*, vol. 25, no. 5, pp. 573–583, 2012.

[124] T. Z. Tan, Q. H. Miow, Y. Miki, T. Noda, S. Mori, R. Y.-J. Huang, and J. P. Thiery, "Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients," *EMBO molecular medicine*, vol. 6, no. 10, pp. 1279–1293, 2014.

[125] F. Garrido, F. Ruiz-Cabello, and N. Aptsiauri, "Rejection versus escape: the tumor mhc dilemma," *Cancer Immunology, Immunotherapy*, vol. 66, no. 2, pp. 259–271, 2017.

[126] D. M. Pardoll, "The blockade of immune checkpoints in cancer immunotherapy," *Nature Reviews Cancer*, vol. 12, no. 4, pp. 252–264, 2012.

[127] J. Gong, A. Chehrazi-Raffle, S. Reddi, and R. Salgia, "Development of pd-1 and pd-l1 inhibitors as a form of cancer immunotherapy: a comprehensive review of registration trials and future considerations," *Journal for immunotherapy of cancer*, vol. 6, no. 1, p. 8, 2018.

[128] S. A. Mani, W. Guo, M.-J. Liao, E. N. Eaton, A. Ayyanan, A. Y. Zhou, M. Brooks, F. Reinhard, C. C. Zhang, M. Shipitsin *et al.*, "The epithelial-mesenchymal transition generates cells with properties of stem cells," *Cell*, vol. 133, no. 4, pp. 704–715, 2008.

[129] Y. Lee, G. Guan, and A. A. Bhagat, "Clearcell® fx, a label-free microfluidics technology for enrichment of viable circulating tumor cells," *Cytometry Part A*, vol. 93, no. 12, pp. 1251–1254, 2018.

[130] N. Ramalingam, B. Fowler, L. Szpankowski, A. A. Leyrat, K. Hukari, M. T. Maung, W. Yorza, M. Norris, C. Cesar, J. Shuga *et al.*, "Fluidic logic used in a systems approach to enable integrated single-cell functional analysis," *Frontiers in bioengineering and biotechnology*, vol. 4, p. 70, 2017.

[131] J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu *et al.*, "Supervised risk predictor of

breast cancer based on intrinsic subtypes," *Journal of clinical oncology*, vol. 27, no. 8, p. 1160, 2009.

[132] D. Bausch-Fluck, A. Hofmann, T. Bock, A. P. Frei, F. Cerciello, A. Jacobs, H. Moest, U. Omasits, R. L. Gundry, C. Yoon *et al.*, "A mass spectrometric-derived cell surface protein atlas," *PloS one*, vol. 10, no. 4, p. e0121314, 2015.

[133] W. Y. Khot and M. Y. Nadkar, "The 2019 novel coronavirus outbreak-a global threat," *J Assoc Physicians India*, vol. 68, no. 3, p. 67, 2020.

[134] M. Cascella, M. Rajnik, A. Cuomo, S. C. Dulebohn, and R. Di Napoli, "Features, evaluation and treatment coronavirus (covid-19)," in *Statpearls [internet]*. StatPearls Publishing, 2020.

[135] D. Lewis, "Coronavirus outbreak: what's next?" *Nature*, vol. 578, no. 7793, pp. 15–16, 2020.

[136] M. Fadel, J. Salomon, and A. Descatha, "Coronavirus outbreak: the role of companies in preparedness and responses," *The Lancet Public Health*, vol. 5, no. 4, p. e193, 2020.

[137] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu *et al.*, "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding," *The Lancet*, vol. 395, no. 10224, pp. 565–574, 2020.

[138] D. Paraskevis, E. G. Kostaki, G. Magiorkinis, G. Panayiotakopoulos, G. Sourvinos, and S. Tsiodras, "Full-genome evolutionary analysis of

the novel corona virus (2019-ncov) rejects the hypothesis of emergence as a result of a recent recombination event," *Infection, Genetics and Evolution*, vol. 79, p. 104212, 2020.

[139] Y. Chen, Q. Liu, and D. Guo, "Emerging coronaviruses: genome structure, replication, and pathogenesis," *Journal of medical virology*, vol. 92, no. 4, pp. 418–423, 2020.

[140] J. F.-W. Chan, K.-H. Kok, Z. Zhu, H. Chu, K. K.-W. To, S. Yuan, and K.-Y. Yuen, "Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting wuhan," *Emerging microbes & infections*, vol. 9, no. 1, pp. 221–236, 2020.

[141] W. Li, M. J. Moore, N. Vasilieva, J. Sui, S. K. Wong, M. A. Berne, M. Somasundaran, J. L. Sullivan, K. Luzuriaga, T. C. Greenough *et al.*, "Angiotensin-converting enzyme 2 is a functional receptor for the sars coronavirus," *Nature*, vol. 426, no. 6965, pp. 450–454, 2003.

[142] H. Hofmann, K. Pyrc, L. Van Der Hoek, M. Geier, B. Berkhout, and S. Pöhlmann, "Human coronavirus nl63 employs the severe acute respiratory syndrome coronavirus receptor for cellular entry," *Proceedings of the National Academy of Sciences*, vol. 102, no. 22, pp. 7988–7993, 2005.

[143] R. Yan, Y. Zhang, Y. Li, L. Xia, Y. Guo, and Q. Zhou, "Structural basis for the recognition of sars-cov-2 by full-length human ace2," *Science*, vol. 367, no. 6485, pp. 1444–1448, 2020.

[144] D. Wrapp, N. Wang, K. S. Corbett, J. A. Goldsmith, C.-L. Hsieh, O. Abiona, B. S. Graham, and J. S. McLellan, "Cryo-em structure of the 2019-ncov spike in the prefusion conformation," *Science*, vol. 367, no. 6483, pp. 1260–1263, 2020.

[145] J. Li, X. Zhou, Y. Zhang, F. Zhong, C. Lin, P. J. McCormick, F. Jiang, H. Zhou, Q. Wang, J. Duan *et al.*, "Crystal structure of sars-cov-2 main protease in complex with a chinese herb inhibitor shikonin," *bioRxiv*, 2020.

[146] A. Hasan, B. A. Paray, A. Hussain, F. A. Qadir, F. Attar, F. M. Aziz, M. Sharifi, H. Derakhshankhah, B. Rasti, M. Mehrabi *et al.*, "A review on the cleavage priming of the spike protein on coronavirus by angiotensin-converting enzyme-2 and furin," *Journal of Biomolecular Structure and Dynamics*, pp. 1–9, 2020.

[147] K. Wang, W. Chen, Y.-S. Zhou, J.-Q. Lian, Z. Zhang, P. Du, L. Gong, Y. Zhang, H.-Y. Cui, J.-J. Geng *et al.*, "Sars-cov-2 invades host cells via a novel route: Cd147-spike protein," *BioRxiv*, 2020.

[148] S. Zaim, J. H. Chong, V. Sankaranarayanan, and A. Harky, "Covid-19 and multi-organ response," *Current Problems in Cardiology*, p. 100618, 2020.

[149] C. Menni, A. Valdes, M. B. Freydin, S. Ganesh, J. E.-S. Moustafa, A. Visconti, P. Hysi, R. C. Bowyer, M. Mangino, M. Falchi *et al.*, "Loss of smell and taste in combination with other symptoms is a strong predictor of covid-19 infection," *MedRxiv*, 2020.

[150] C. H. Yan, F. Faraji, D. P. Prajapati, B. T. Ostrander, and A. S. DeConde, "Self-reported olfactory loss associates with outpatient clinical course in covid-19," in *International Forum of Allergy & Rhinology*. Wiley Online Library, 2020.

[151] J. R. Lechien, P. Cabaraux, C. Chiesa-Estomba, M. Khalife, J. Plzak, S. Hans *et al.*, "Objective olfactory testing in patients presenting with sudden onset olfactory dysfunction as the first manifestation of confirmed covid-19 infection," *Medrxiv*, 2020.

[152] C. R. Chen, C. Kachramanoglou, D. Li, P. Andrews, and D. Choi, "Anatomy and cellular constituents of the human olfactory mucosa: a review," *Journal of Neurological Surgery Part B: Skull Base*, vol. 75, no. 05, pp. 293–300, 2014.

[153] G. Ahuja, S. B. Nia, V. Zapilko, V. Shiriagin, D. Kowatschew, Y. Oka, and S. I. Korsching, "Kappe neurons, a novel population of olfactory sensory neurons," *Scientific reports*, vol. 4, no. 1, pp. 1–8, 2014.

[154] B. Malnic, P. A. Godfrey, and L. B. Buck, "The human olfactory receptor gene family," *Proceedings of the National Academy of Sciences*, vol. 101, no. 8, pp. 2584–2589, 2004.

[155] F. Liang, "Sustentacular cell enwrapment of olfactory receptor neuronal dendrites: An update," *Genes*, vol. 11, no. 5, p. 493, 2020.

[156] D. T. Moran, J. C. Rowley, B. W. Jafek, and M. A. Lovell, "The fine structure of the olfactory mucosa in man," *Journal of neurocytology*, vol. 11, no. 5, pp. 721–746, 1982.

[157] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija, "Comprehensive integration of single-cell data," *Cell*, vol. 177, no. 7, pp. 1888–1902, 2019.

[158] A. M. Joiner, W. W. Green, J. C. McIntyre, B. L. Allen, J. E. Schwob, and J. R. Martens, "Primary cilia on horizontal basal cells regulate regeneration of the olfactory epithelium," *Journal of Neuroscience*, vol. 35, no. 40, pp. 13 761–13 772, 2015.

[159] J. E. Schwob, W. Jang, E. H. Holbrook, B. Lin, D. B. Herrick, J. N. Peterson, and J. Hewitt Coleman, "Stem and progenitor cells of the mammalian olfactory epithelium: Taking poietic license," *Journal of Comparative Neurology*, vol. 525, no. 4, pp. 1034–1054, 2017.

[160] M. A. Durante, S. Kurtenbach, Z. B. Sargi, J. W. Harbour, R. Choi, S. Kurtenbach, G. M. Goss, H. Matsunami, and B. J. Goldstein, "Single-cell analysis of olfactory neurogenesis and differentiation in adult humans," *Nature neuroscience*, vol. 23, no. 3, pp. 323–326, 2020.

[161] D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, K. M. White, M. J. O'Meara, V. V. Rezelj, J. Z. Guo, D. L. Swaney *et al.*, "A sars-cov-2 protein interaction map reveals targets for drug repurposing," *Nature*, pp. 1–13, 2020.

[162] L. R. Saraiva, F. Riveros-McKay, M. Mezzavilla, E. H. Abou-Moussa, C. J. Arayata, M. Makhlouf, C. Trimmer, X. Ibarra-Soria, M. Khan, L. Van Gerven *et al.*, "A transcriptomic atlas of mammalian olfactory

mucosae reveals an evolutionary influence on food odor detection in humans," *Science Advances*, vol. 5, no. 7, p. eaax0396, 2019.

[163] A. Dereeper, V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J.-F. Dufayard, S. Guindon, V. Lefort, M. Lescot *et al.*, "Phylogeny. fr: robust phylogenetic analysis for the non-specialist," *Nucleic acids research*, vol. 36, no. suppl_2, pp. W465–W469, 2008.

[164] S. Whelan and N. Goldman, "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach," *Molecular biology and evolution*, vol. 18, no. 5, pp. 691–699, 2001.

[165] F. Chevenet, C. Brun, A.-L. Bañuls, B. Jacq, and R. Christen, "Treedyn: towards dynamic graphics and annotations for analyses of trees," *BMC bioinformatics*, vol. 7, no. 1, p. 439, 2006.

[166] J. Shang, G. Ye, K. Shi, Y. Wan, C. Luo, H. Aihara, Q. Geng, A. Auerbach, and F. Li, "Structural basis of receptor recognition by sars-cov-2," *Nature*, vol. 581, no. 7807, pp. 221–224, 2020.

[167] A. Šali and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints," *Journal of molecular biology*, vol. 234, no. 3, pp. 779–815, 1993.

[168] G. Van Zundert, J. Rodrigues, M. Trellet, C. Schmitz, P. Kastritis, E. Karaca, A. Melquiond, M. van Dijk, S. De Vries, and A. Bonvin, "The haddock2. 2 web server: user-friendly integrative modeling of biomolec-

ular complexes," *Journal of molecular biology*, vol. 428, no. 4, pp. 720–725, 2016.

[169] L. C. Xue, J. P. Rodrigues, P. L. Kastritis, A. M. Bonvin, and A. Vangone, "Prodigy: a web server for predicting the binding affinity of protein–protein complexes," *Bioinformatics*, vol. 32, no. 23, pp. 3676–3678, 2016.

[170] F. Sievers and D. G. Higgins, "Clustal omega, accurate alignment of very large numbers of sequences," in *Multiple sequence alignment methods*. Springer, 2014, pp. 105–116.

[171] P. Padmanabhan, R. Desikan, and N. Dixit, "Targeting tmprss2 and cathepsin b/l together may be synergistic against sars-cov-2 infection," 2020.

[172] M. H. Naghavi and D. Walsh, "Microtubule regulation and function during virus infection," *Journal of virology*, vol. 91, no. 16, 2017.

[173] N. Van Doremalen, T. Bushmaker, D. H. Morris, M. G. Holbrook, A. Gamble, B. N. Williamson, A. Tamin, J. L. Harcourt, N. J. Thornburg, S. I. Gerber *et al.*, "Aerosol and surface stability of sars-cov-2 as compared with sars-cov-1," *New England Journal of Medicine*, vol. 382, no. 16, pp. 1564–1567, 2020.

[174] C. Shan, Y.-F. Yao, X.-L. Yang, Y.-W. Zhou, J. Wu, G. Gao, Y. Peng, L. Yang, X. Hu, J. Xiong *et al.*, "Infection with novel coronavirus (sars-cov-2) causes pneumonia in the rhesus macaques," 2020.

[175] W. Tai, L. He, X. Zhang, J. Pu, D. Voronin, S. Jiang, Y. Zhou, and L. Du, "Characterization of the receptor-binding domain (rbd) of 2019 novel coronavirus: implication for development of rbd protein as a viral attachment inhibitor and vaccine," *Cellular & molecular immunology*, vol. 17, no. 6, pp. 613–620, 2020.

[176] G. Chen, B. Ning, and T. Shi, "Single-cell rna-seq technologies and related computational data analysis," *Frontiers in genetics*, vol. 10, p. 317, 2019.

[177] M. L. Getchell and T. V. Getchell, "Fine structural aspects of secretion and extrinsic innervation in the olfactory mucosa," *Microscopy research and technique*, vol. 23, no. 2, pp. 111–127, 1992.

[178] R. L. Doty, *Handbook of olfaction and gustation*. John Wiley & Sons, 2015.

[179] L. H. Bannister and H. C. Dodson, "Endocytic pathways in the olfactory and vomeronasal epithelia of the mouse: ultrastructure and uptake of tracers," *Microscopy research and technique*, vol. 23, no. 2, pp. 128–141, 1992.

[180] A. R. Dahl, W. M. Hadley, F. F. Hahn, J. M. Benson, and R. O. McClellan, "Cytochrome p-450-dependent monooxygenases in olfactory epithelium of dogs: possible role in tumorigenicity," *Science*, vol. 216, no. 4541, pp. 57–59, 1982.

[181] Y. Suzuki, M. Takeda, and A. I. Farbman, "Supporting cells as phagocytes in the olfactory epithelium after bulbectomy," *Journal of Comparative Neurology*, vol. 376, no. 4, pp. 509–517, 1996.

[182] W. Breipohl, H. Laugwitz, and N. Bornfeld, "Topological relations between the dendrites of olfactory sensory cells and sustentacular cells in different vertebrates. an ultrastructural study." *Journal of anatomy*, vol. 117, no. Pt 1, p. 89, 1974.

[183] J. A. Rafols and T. V. Getchell, "Morphological relations between the receptor neurons, sustentacular cells and schwann cells in the olfactory mucosa of the salamander," *The Anatomical Record*, vol. 206, no. 1, pp. 87–101, 1983.

[184] C. C. Hegg, M. Irwin, and M. T. Lucero, "Calcium store-mediated signaling in sustentacular cells of the mouse olfactory epithelium," *Glia*, vol. 57, no. 6, pp. 634–644, 2009.

[185] N. Iwai, Z. Zhou, D. R. Roop, and R. R. Behringer, "Horizontal basal cells are multipotent progenitors in normal and injured adult olfactory epithelium," *Stem cells*, vol. 26, no. 5, pp. 1298–1306, 2008.

[186] D. B. Herrick, B. Lin, J. Peterson, N. Schnittke, and J. E. Schwob, "Notch1 maintains dormancy of olfactory horizontal basal cells, a reserve neural stem cell," *Proceedings of the National Academy of Sciences*, vol. 114, no. 28, pp. E5589–E5598, 2017.

[187] S. Zhu, T. Qing, Y. Zheng, L. Jin, and L. Shi, "Advances in single-cell rna sequencing and its applications in cancer research," *Oncotarget*, vol. 8, no. 32, p. 53763, 2017.

[188] P. Kumar, Y. Tan, and P. Cahan, "Understanding development and stem cells using single cell-based analyses of gene expression," *Development*, vol. 144, no. 1, pp. 17–32, 2017.

[189] D. Sengupta, N. A. Rayan, M. Lim, B. Lim, and S. Prabhakar, "Fast, scalable and accurate differential expression analysis for single cells," *BioRxiv*, p. 049734, 2016.

[190] G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic *et al.*, "Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data," *Genome biology*, vol. 16, no. 1, p. 278, 2015.

[191] T. N. Vu, Q. F. Wills, K. R. Kalari, N. Niu, L. Wang, M. Rantalainen, and Y. Pawitan, "Beta-poisson model for single-cell rna-seq data analyses," *Bioinformatics*, vol. 32, no. 14, pp. 2128–2135, 2016.

[192] G. Martínez-Mekler, R. A. Martínez, M. B. del Río, R. Mansilla, P. Miramontes, and G. Cocho, "Universality of rank-ordering distributions in the arts and sciences," *PLoS One*, vol. 4, no. 3, 2009.

[193] 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, "Pseudo-

temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions," *Nature biotechnology*, vol. 32, no. 4, p. 381, 2014.

[194] A. Iyer, K. Gupta, S. Sharma, K. Hari, Y. F. Lee, N. Ramalingam, Y. S. Yap, J. West, A. A. Bhagat, B. V. Subramani *et al.*, "Integrative analysis and machine learning based characterization of single circulating tumor cells," *Journal of clinical medicine*, vol. 9, no. 4, p. 1206, 2020.

[195] C. Soneson and M. D. Robinson, "Bias, robustness and scalability in single-cell differential expression analysis," *Nature methods*, vol. 15, no. 4, p. 255, 2018.

[196] M. Love, S. Anders, and W. Huber, "Differential analysis of count data–the deseq2 package," *Genome Biol*, vol. 15, no. 550, pp. 10–1186, 2014.

[197] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, "voom: Precision weights unlock linear model analysis tools for rna-seq read counts," *Genome biology*, vol. 15, no. 2, p. R29, 2014.

[198] G. Casella and R. Berger, "Statistical inference . pacific grove, ca: Thomson learning," 2002.

[199] I. J. Myung, "Tutorial on maximum likelihood estimation," *Journal of mathematical Psychology*, vol. 47, no. 1, pp. 90–100, 2003.

[200] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Nature Precedings*, pp. 1–1, 2010.

[201] A. Giustacchini, S. Thongjuea, N. Barkas, P. S. Woll, B. J. Povinelli, C. A. Booth, P. Sopp, R. Norfo, A. Rodriguez-Meira, N. Ashley *et al.*, "Single-

cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia," *Nature medicine*, vol. 23, no. 6, p. 692, 2017.

[202] S. N. Hart, T. M. Therneau, Y. Zhang, G. A. Poland, and J.-P. Kocher, "Calculating sample size estimates for rna sequencing data," *Journal of computational biology*, vol. 20, no. 12, pp. 970–978, 2013.

[203] 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, p. 411, 2018.

[204] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "Rocr: visualizing classifier performance in r," *Bioinformatics*, vol. 21, no. 20, pp. 3940–3941, 2005.

[205] T. O. Kvålseth, "Note on cohen's kappa," *Psychological reports*, vol. 65, no. 1, pp. 223–226, 1989.

[206] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, "proc: an open-source package for r and s+ to analyze and compare roc curves," *BMC bioinformatics*, vol. 12, no. 1, pp. 1–8, 2011.