Machine learning approaches in cancer detection and treatment

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

SARITA
(PhD17201)

Department of Computational Biology
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI
NEW DELHI- 110020

December 2023

THESIS CERTIFICATE

This is to certify that the thesis titled **Machine learning approaches in cancer detection and treatment**, submitted by **Sarita**, to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of **Doctor of Philosophy**, is a bona fide record of the research work done by her under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Debarka Sengupta
Department of Computational Biology
Indraprastha Institute of Information Technology (IIIT-Delhi)
New Delhi, India, 110020

Signature: 
Date: 12/27/2023

Prof. Lalit Kumar
Retd. Head Department of Medical Oncology
All India Institute Of Medical Science (AIIMS-Delhi)
New Delhi, India, 110029

Signature: 
Date: 12/28/2023

Place: New Delhi

ACKNOWLEDGEMENTS

I express my deepest gratitude to my esteemed supervisor, Dr. Debarka Sengupta, for providing me with the guidance and counsel I needed to succeed in the Ph.D. program. He has been an incredible source of inspiration, understanding, and mentorship throughout my journey. Dr. Sengupta has been instrumental in mapping my Ph.D. journey, advising on a research topic, connecting me with the needed resources, and always being available to respond to my emails and questions. He values every step I take and celebrates every accomplishment I make. His comfortable and friendly work environment has led to fruitful collaborations, and his way of helping humanity is truly admirable. I will always be indebted to him for his unwavering support and patience. I could not have imagined having a better advisor and mentor for my Ph.D. study.

In addition to my supervisor, I am incredibly grateful to my co-supervisor, Dr. Lalit Kumar, who provided novel data for one of my thesis's significant projects. I appreciate his consistent motivation, patience, vast knowledge, and professional advice for the success of this project.

I sincerely thank Dr. Naveen Ramalingam and Dr. Angshul Majumdar for their research collaboration and for co-mentoring me. Their guidance and support were invaluable, and their contribution to novel and valuable data helped me apply computational approaches to aid in societal benefits.

I thank the University Grants Commission for financially supporting my Ph.D. fellowship. I am also thankful to all the Department of Computational Biology faculty members for their constructive support and the valuable learnings they have shared in the courses they taught. I appreciate Ms. Priti Patel and other administrative staff members' prompt administrative support whenever I needed it. I thank the IT department, especially Mr. Adarsh Agarwal, for their technical assistance. I am also grateful to Anurag Goel for his help with computational methods. Finally, I would like to thank all my colleagues, collaborators, and friends, without whom this journey would not have been possible.

I had like to express my appreciation for my friends, juniors, and colleagues, Chitrita, Namrata, Vivek, Sumeet, Shruti, Samriddhi, Krishan, Anjali Dhall, Neelam, Chakit, Dilraj, Shiju, Naina, Taniya, Bernadette, Tanya, Urvashi, Abhishek, Himani for creating a friendly and supportive work environment. I am grateful for the role played by some of my closest friends, Anjali Lathwal, Neetesh, Smriti, Priyadarshini, Omkar, Shreya, Raghav, Madhu, Ishant, and Indra Prakash Jha, who were pillars of support during this journey.

I appreciate my best friend, Vinita Lamba, my siblings Anita, Kavita, Vijay, and Sonu, and my sister-in-law Nisham for being a great source of strength and support. I would like to express my sincere gratitude towards my mother-in-law, Reshma Devi, for her unwavering love, moral support, and tremendous help in supporting my work. Thanks to my father-in-law, Raj Singh, for his constant encouragement and support.

I owe a debt of gratitude to an extraordinary person, my husband, Kulvir Singh, for his continued and unfailing love, support, and understanding during my pursuit of a Ph.D. Completing my thesis would not have been possible without his unwavering presence in my life. Sometimes I felt like giving up, but his guidance and support helped me stay focused and motivated. I am deeply grateful for his belief in me and his unwavering commitment to our family. I am also thankful to our little girl Mapril (Veera) for her patience and understanding during this journey. I am blessed to have such a lovely and caring family standing beside me with their love and unconditional support.

Finally, I would like to acknowledge my parents, Maa Bimla Devi and Pappa Ram Singh, for showing faith in me and giving me the liberty to choose what I desire. Your selfless love, care, pain, and sacrifice have shaped my life in more ways than I can express. Although you hardly understood what I researched, you were always willing to support my decision. I would never be able to repay the love and affection showered upon me by my parents.

Last, I thank the Almighty for giving me the strength and patience to work through all these years. Without His blessings, I would not be standing here today with my head held high. Thank you all for participating in my journey and contributing to my success.

Sarita

List of publications

Publications

1. **Poonia S**, Goel A, Chawla S, Bhattacharya N, Rai P, Lee YF, Yap YS, West J, Bhagat AA, Tayal J, Mehta A, Ahuja G, Majumdar A, Ramalingam N, Sengupta D. Marker-free characterization of full-length transcriptomes of single live circulating tumor cells. Genome Res. 2023 Jan;33(1):80-95. doi: 10.1101/gr.276600.122. Epub 2022 Nov 22. PMID: 36414416; PMCID: PMC9977151.

2. Goswami C, **Poonia S**, Kumar L, Sengupta D. Staging System to Predict the Risk of Relapse in Multiple Myeloma Patients Undergoing Autologous Stem Cell Transplantation. Front Oncol. 2019 Jul 12;9:633. doi: 10.3389/fonc.2019.00633. PMID: 31355145; PMCID: PMC6640159. **(* Co-first author)**

3. **Poonia, S.**, Chawla, S., Kaushik, S. R., & Sengupta, D. (2019). Pathway Informatics. Encyclopedia of Bioinformatics and Computational Biology. 2019;796-804, doi: https://doi.org/10.1016/B978-0-12-809633-8.20288-4. ISBN: 978-0-12-811432-2.

Other publications

1. Chawla S, Rockstroh A, Lehman M, Ratther E, Jain A, Anand A, Gupta A, Bhattacharya N, **Poonia S**, Rai P, Das N, Majumdar A, Jayadeva, Ahuja G, Hollier BG, Nelson CC, Sengupta D. Gene expression based inference of cancer drug sensitivity. Nat Commun. 2022 Sep 27;13(1):5680. doi: 10.1038/s41467-022-33291-z. PMID: 36167836; PMCID: PMC9515171.

2. Rai P, Jain A, Jha N, Sharma D, Kumar S, Raj A, Gupta A, **Poonia S**, Chawla S, Majumdar A, Chakraborty T. A visual atlas of genes' tissue-specific pathological roles. bioRxiv. 2022 Jan 10:2022-01. doi: https://doi.org/10.1101/2022.01.08.475476

ABSTRACT

Cancer has become the second leading cause of mortality worldwide, and early detection and adequate treatment are crucial in reducing the cancer burden. Metastasis, which involves malignant cells detaching from the primary tumor and colonizing other distant organs, is the leading cause of cancer-related deaths. The microenvironment, immune cells, stromal cells, and drug selection pressures influence tumors' heterogeneity and dynamicity, making it challenging to select the most effective treatment approach throughout the entire course of the disease. Liquid biopsy and single-cell transcriptomics have emerged as promising techniques for cancer detection. Bodily fluids such as blood, urine, and saliva provide rich biomarkers. Circulating tumor cells and other tumor-associated cell products have been identified in the bloodstream, providing potential biomarkers for cancer detection. Through serial blood analysis, liquid biopsy techniques can help track spatial and temporal heterogeneity in tumor biology. Characterizing circulating tumor cells (CTCs) provides essential biological information about the disease as they are the primary live tumor cells responsible for metastasis. Existing CTC detection methods rely on surface markers, which may be shed during the epithelial-to-mesenchymal (EMT) process or due to various stressors in the blood. Therefore, marker-free detection and characterization of CTCs are necessary. To achieve the best possible outcomes, it is crucial to manage cancer and any clinical factors that may impact treatment response or contribute to disease relapse. By identifying and addressing these factors, healthcare providers can develop effective treatment plans and improve overall cancer management. This approach can help patients achieve longer-term remission and better quality of life.

Over the past two decades, machine learning (ML) has shown tremendous potential in enhancing cancer diagnosis and treatment accuracy and efficiency. Our research leverages the power of ML to address the pressing need for timely cancer detection and optimal management of the disease. By employing advanced ML algorithms, we aimed to improve the accuracy and speed of cancer diagnosis, identify the most effective treatment options, and enable personalized cancer care.

For marker-free detection and characterization of CTCs, we created a novel unsupervised clustering algorithm, unCTC, which can leverage single-cell transcriptomic data to detect and characterize CTCs. In unCTC, a wide range of computational and statistical modules are integrated, such as novel Deep Dictionary Learning with *k-means* Clustering Cost (DDLK) approach for scRNA-Seq clustering, expression-based inference of copy number variation (CNV), and combinatorial, marker-based validation of malignant phenotypes. DDLK provides a robust separation of circulating tumor cells (CTCs) and white blood cells (WBCs) in the pathway space, unlike the gene expression space. The utility of unCTC was validated on single-cell RNA sequencing (scRNA-Seq) profiles of breast CTCs from six patients. These CTCs were profiled using an integrated ClearCell® FX and Polaris™ workflow that relies on the size-based separation of CTCs and marker-based WBC depletion.

Apart from detecting and treating cancer, it is of utmost importance to effectively manage various clinical factors that play a role in disease relapse. To investigate this, we conducted an analysis using clinical data from 253 patients who received Autologous Stem Cell Transplantation (ASCT) therapy at All India Institute for Medical Sciences (AIIMS) in Delhi. We employed a combination of Boruta, random forest, and Bayesian network analysis methods to identify important factors contributing to relapse or progression-free survival of multiple myeloma patients who underwent autologous stem cell transplantation. We have identified five multivariate factors that significantly influence the success of ASCT treatment. Additionally, we trained a random forest model using the top 5 attributes that significantly impact ASCT treatment relapse, which can classify patients into two categories: ASCT relapse within 30 months or not, with an accuracy of 0.76.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Cancer is a disease with multiple factors that develops from the accumulation of genetic and epigenetic changes in cells (Baylin and Jones, 2016; Takeshima and Ushijima, 2019). The World Health Organization (WHO) defines cancer as the abnormal growth and spread of cells that can invade and damage the healthy tissues around them (National Institutes of Health (US) and Biological Sciences Curriculum Study, 2007). There are many types and subtypes of cancer, which can be broadly categorized into five major types: carcinomas, sarcomas, leukemias, lymphomas, and central nervous system (CNS) cancers (Tarver, 2012). Carcinomas comprise the majority of cancer cases, comprising around 80% of all diagnosed cancers (Sung *et al.*, 2021). Carcinomas develop from the epithelial cells that form the lining of both internal and external surfaces, such as the skin, lungs, breast, prostate, and colon. Conversely, rare Sarcomas arise from connective tissues, such as bone, muscle, and cartilage. Leukemias and lymphomas arise from immune system cells, with leukemias affecting the blood and bone marrow and lymphomas affecting the lymphatic system. CNS cancers are a heterogeneous group of tumors that arise in the brain and spinal cord (Louis *et al.*, 2021).

Cancer prevalence and incidence vary depending on various factors, including age, gender, race, lifestyle, and environmental exposures. The American Cancer Society (2022) reported that around 1.9 million new cancer cases were identified in the United States during 2021, with an estimated 609,360 deaths resulting from cancer. According to the WHO (2021), cancer caused approximately 9.6 million deaths in 2018, making it the second most prevalent cause of death worldwide. Additionally, cancer is a significant global health concern, resulting in nearly 10 million deaths in 2020. The most common new cancer cases in 2020 were breast, lung, colon and rectum, prostate, skin (non-melanoma), and stomach cancer. In 2020, breast, lung, colon and rectum, prostate, non-melanoma skin, and stomach cancers were the most frequently diagnosed new cancer cases. The primary contributors to cancer-related deaths in 2020 were lung, colon and rectum, liver, stomach, and breast cancers. Yearly, roughly 400,000 children are

diagnosed with cancer; the prevalent types differ from country to country (Ferlay *et al.*, 2021).

Early detection and treatment of cancer can prevent up to one-third of cancer-related deaths. Suitable prevention, detection, and treatment strategies can save millions of lives each year by prioritizing cancer prevention and control efforts. Encouraging healthy lifestyle choices such as a nutritious diet, avoiding tobacco and alcohol, and staying physically active can also reduce cancer risk.

1.1 Hallmarks of cancer

The Hallmarks of Cancer encompass a set of six crucial attributes that characterize cancer cells. These defining features have gained widespread acceptance within the scientific community and serve as a fundamental framework for comprehending the intricacies of cancer biology and devising innovative therapies **(Figure 1.1)** (Hanahan and Weinberg, 2011).

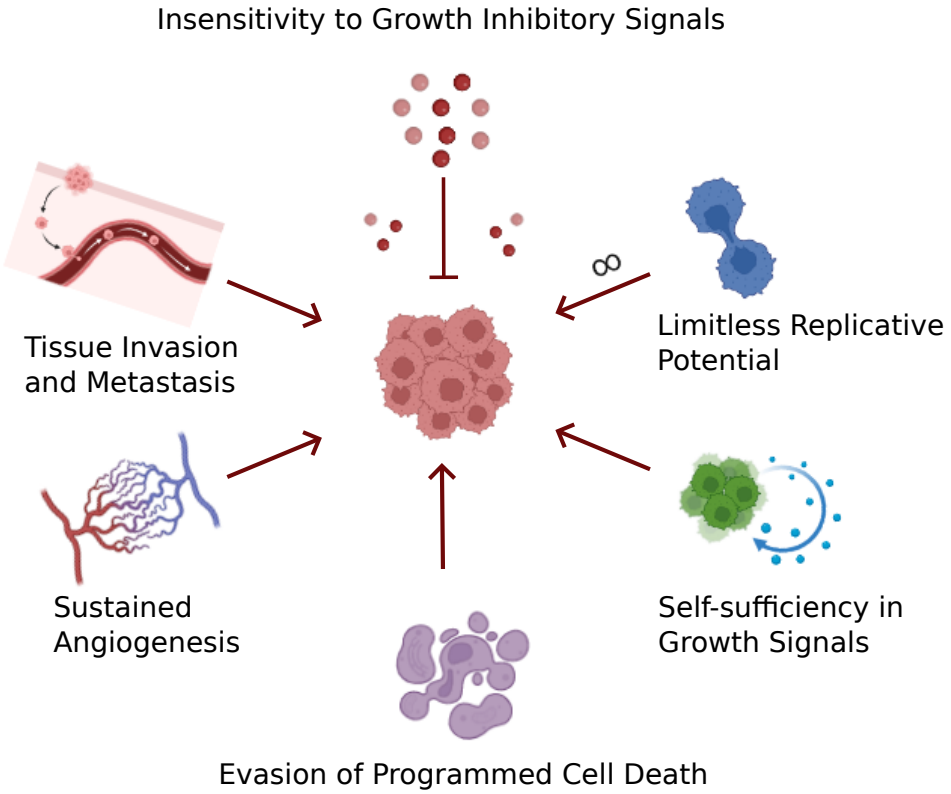

Figure 1.1: Hallmarks of cancer

1.1.1 Self-sufficiency in growth signals

Cancerous cells exhibit the ability to proliferate autonomously, independent of standard growth signals. They have gene mutations that control cell growth and division, activating signaling pathways that drive uncontrolled cell proliferation (Hanahan and Weinberg, 2000, 2011).

1.1.2 Insensitivity to inhibitory growth signals

Cancerous cells can circumvent signals that inhibit cellular growth. They can become resistant to the effects of growth-inhibitory signals, such as TGF-β, which would typically trigger cell cycle arrest (Hanahan and Weinberg, 2000; Gutschner and Diederichs, 2012). They can become resistant to the effects of growth-inhibitory signals, such as TGF-β, which would typically trigger cell cycle arrest (Hanahan and Weinberg, 2000; Gutschner and Diederichs, 2012).

1.1.3 Evasion of programmed cell death (Apoptosis)

Cancer cells can avoid programmed cell death, or apoptosis, a natural process that eliminates damaged or abnormal cells. They have mutations in genes that control apoptosis, allowing them to survive and continue to divide (Hanahan and Weinberg, 2000; Gutschner and Diederichs, 2012; Chen *et al.*, 2018).

1.1.4 Limitless replicative potential

Unlike normal cells with a finite lifespan, cancerous cells can undergo unlimited cell division. Cancerous cells accomplish this feat via the activation of telomerase, an enzyme that preserves the length of telomeres. Telomeres are protective caps at the ends of chromosomes that experience shortening with each successive cell division (Hanahan and Weinberg, 2011; Gutschner and Diederichs, 2012).

1.1.5 Sustained angiogenesis

Cancerous cells can stimulate angiogenesis, forming new blood vessels to obtain the necessary nutrients and oxygen required for their proliferation and sustenance. They secrete growth factors that promote angiogenesis and can also stimulate the surrounding tissue to release angiogenic factors (Hu *et al.*, 2015; Al-Ostoot *et al.*, 2021).

1.1.6 Tissue invasion and metastasis

The process leading to cancer development is intricate and involves the acquisition of distinct features by malignant cells, including uncontrolled cell growth and the ability to invade and metastasize into nearby tissues and organs (Martin *et al.*, 2013). Metastases of tumor cells cause up to 90% of cancer-related deaths in distant organs (Seyfried and Huysentruyt, 2013). Cancer cells can spread and settle in new areas of the body that offer them favorable conditions to thrive. Cancer progression is a complex and stepwise process known as the invasion-metastasis cascade. This sequence of events involves multiple alterations in the biology of cancer cells, commencing with their local invasion and progressing through the subsequent stages of tumor cell intravasation, movement via the lymphatic system, extravasation, micrometastases, and finally, the colonization of new sites (Martin *et al.*, 2013; van Zijl *et al.*, 2011).

1.2 Cancer diagnosis and detection

1.2.1 Traditional cancer diagnostic methods

Traditional cancer diagnostic methods refer to the standard techniques doctors have used for many years to identify the presence of cancer in a patient. These methods include physical exams, laboratory tests, imaging tests, and biopsy procedures (Soper and Rasooly, 2016).

1.2.1.1 Physical exam

A physical exam involves a doctor examining a patient's body for any signs of cancer, such as lumps or swelling. The doctor will also ask the patient about their medical history and any symptoms they may be experiencing (Institute and National Cancer Institute, 2020; Soper and Rasooly, 2016).

1.2.1.2 Laboratory test

Laboratory tests are used to analyze blood or urine samples for abnormal cells or biomarkers that could indicate the presence of cancer (McPherson and Pincus, 2017). Elevated levels of white blood cells may serve as a potential indicator of leukemia, a malignant neoplasm affecting hematopoietic stem cells in the bone marrow. This leads to an abnormal increase in immature or abnormal white blood cell production (Soper and Rasooly, 2016).

1.2.1.3 Imaging test

Medical imaging procedures encompass a variety of methodologies, such as computed tomography (CT), magnetic resonance imaging (MRI), and X-ray imaging, that employ different physical principles to capture high-resolution visual representations of the internal structures of the human body (Prince and Links, 2008). These images can reveal the presence of tumors or other abnormal growths that may be cancerous (Institute and National Cancer Institute, 2020; Soper and Rasooly, 2016).

1.2.1.4 Biopsy procedures

Biopsy procedures are a standard diagnostic method used to confirm the presence of cancerous cells in a patient's body (Eloubeidi *et al.*, 2003). A biopsy procedure involves the extraction of a minute quantity of tissue or fluid from a region of the body exhibiting anomalous pathological changes, which is subsequently subjected to microscopic examination to detect the presence or absence of cancerous cells (Soper and Rasooly, 2016; Schnitt and Collins, 2009).

Various types of biopsy procedures are available, including.

- **Fine-needle aspiration biopsy (FNAB):** During this minimally invasive procedure, a thin needle is inserted into the tumor or mass to extract cells for examination. This procedure is typically used for smaller masses or those close to the skin's surface (Chaiwun and Thorner, 2007; Chuo and Corder, 2003; He *et al.*, 2007).
- **Core needle biopsy (CNB):** A larger needle extracts a small cylinder of tissue from the tumor or mass. This procedure is often used for larger masses or those that are deeper within the body (Reynolds, 2000; Rosen *et al.*, 2002; Lan *et al.*, 2020).
- **Vacuum-assisted biopsy (VAB):** It is a variation of core needle biopsy that involves using a larger needle to extract multiple tissue samples with the assistance of vacuum pressure. This procedure can be particularly useful for larger masses or those that have irregular shapes (Park and Hong, 2014; Nakano *et al.*, 2018).
- **Incisional biopsy:** A small incision is made in the skin to remove a piece of the tumor or mass for examination. This type of biopsy is typically used when the tumor or mass is too large to be removed entirely (Golden and Hooley, 1994; Parker and Israel, 1996; Heywang-Köbrunner *et al.*, 1999).
- **Excisional biopsy:** Excisional biopsy involves removing the entire tumor or mass for examination. This type of biopsy is often used when the tumor or mass is small and easily accessible (Golden and Hooley, 1994; Parker and Israel, 1996; Heywang-Köbrunner *et al.*, 1999).
- **Endoscopic biopsy:** A thin, flexible tube with a light and camera on end is used to examine and biopsy tissue in areas of the body that are difficult to access, such as the digestive tract or lungs (Levine *et al.*, 1993; Reid *et al.*, 2000; Vaiphei, 2021).

The selection of a particular type of biopsy is influenced by a multitude of factors, including the size and anatomical location of the tumor or lesion, the patient's medical history, and their overall health condition (Pacella *et al.*, 2005). Biopsy procedures are generally safe and well-tolerated, and most patients are able to resume normal activities shortly after the procedure. However, as with any medical procedure, there are potential risks and complications, and patients should discuss these with their healthcare provider before undergoing a biopsy (Elston *et al.*, 2016; Puget *et al.*, 2012; Kelly *et al.*, 2015).

1.2.2 Liquid biopsy: A breakthrough in cancer diagnosis

Liquid biopsy is a revolutionary cancer diagnosis technique involving the detection and analysis of cancer biomarkers in blood, urine, and other body fluids (Poulet *et al.*, 2019; Michela, 2021; Ma and Jeffrey, 2020). The approach relies on cancer cells releasing specific molecules, such as CTCs, miRNA, exosomes, and circulating tumor DNA

(ctDNA), into the bloodstream, which can be isolated and analyzed to detect the presence of cancer and monitor its progression **(Figure 1.2)** (Giannopoulou *et al.*, 2018; Alix-Panabières and Pantel, 2013; Temraz *et al.*, 2022; Jia *et al.*, 2017; Palacín-Aliana *et al.*, 2021). One of the key advantages of liquid biopsy is its non-invasive nature, which eliminates the need for invasive procedures such as tissue biopsies (Marrugo-Ramírez *et al.*, 2018). Moreover, liquid biopsy has the potential to offer up-to-date information about the current state of cancer in an individual and treatment response, enabling personalized and timely interventions. Liquid biopsy is also helpful in monitoring minimal residual disease (MRD) and detecting cancer recurrence, as it can detect the presence of residual cancer cells in patients who have undergone treatment and have no visible signs of cancer (Wang *et al.*, 2021; Lu *et al.*, 2019; Pantel and Alix-Panabières, 2019; Palacín-Aliana *et al.*, 2021).

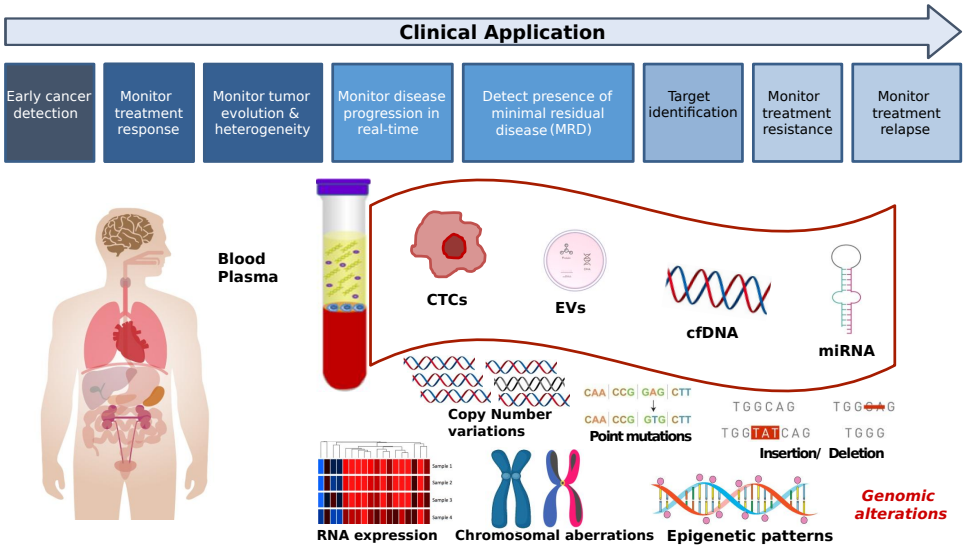

Figure 1.2: Understanding liquid biopsy components. Components of a liquid biopsy derived from peripheral blood, which comprises several tumoral constituents, including micro-RNA (miRNA), circulating cell-free DNA (cfDNA), circulating tumor cells (CTCs), and extracellular vesicles (EVs). These constituents can be separated and analyzed to detect various genomic aberrations specific to the tumor, such as point mutations, copy number variations, structural rearrangements, or epigenetic modifications (Palacín-Aliana *et al.*, 2021).

However, liquid biopsy is still a developing technology, and several challenges must be addressed. For instance, the specificity and sensitivity of liquid biopsy assays need to be improved to reduce false-negative and false-positive results (Ma *et al.*, 2015; Porras *et al.*, 2018). Additionally, the variability of cancer biomarker release and the hetero-

geneity of cancer cells can affect the accuracy and reliability of liquid biopsy results (Porras *et al.*, 2018; Ignatiadis *et al.*, 2021). Liquid biopsy is a promising technique that can transform cancer diagnosis and management. With continued research and development, a liquid biopsy may become a routine clinical tool for cancer detection, monitoring, and treatment (Pacia *et al.*, 2020; Fortunato *et al.*, 2019; Palacín-Aliana *et al.*, 2021).

1.2.3 Circulating tumor cells: A key component of liquid biopsy analysis

CTCs are cells that have dissociated from the primary tumor and entered the bloodstream. They can circulate through the vascular system and eventually settle in a distant location, potentially leading to the development of secondary tumors, known as metastases. CTCs have been identified in the peripheral blood of patients with various types of cancer, and their presence is associated with a worse prognosis (Alix-Panabières and Pantel, 2013; Jin *et al.*, 2019*a*). Due to their rarity and heterogeneity, CTCs are a challenging target for liquid biopsy analysis, but recent advances in technology have enabled their isolation and characterization, allowing for the detection of tumor-specific genomic alterations and the monitoring of disease progression and treatment response (Markou *et al.*, 2011; Hodgkinson *et al.*, 2014; Haber and Velculescu, 2014).

1.2.3.1 Unraveling the mechanisms of cancer progression: Investigating the role of circulating tumor cells

CTCs play a crucial role in the metastatic cascade and are associated with poor clinical outcomes in cancer patients (Serrano and Malapelle, 2023). Understanding the mechanisms underlying CTCs' survival, invasion, and colonization of distant sites is essential for developing effective cancer treatments and preventing disease progression (Lin *et al.*, 2021). Recent studies have investigated the molecular and cellular mechanisms that drive CTCs' phenotypic plasticity and enable them to escape from immune surveillance and survive in circulation. These studies have identified key signaling pathways, such as epithelial-mesenchymal transition (EMT), that regulate CTCs' motility, invasiveness, and stemness properties **(Figure 1.3)** (Samatov *et al.*, 2013). Investigat-

ing the molecular mechanisms that underlie CTCs' behavior and interaction with the microenvironment can provide new insights into the biology of cancer metastasis and guide the development of targeted therapies.

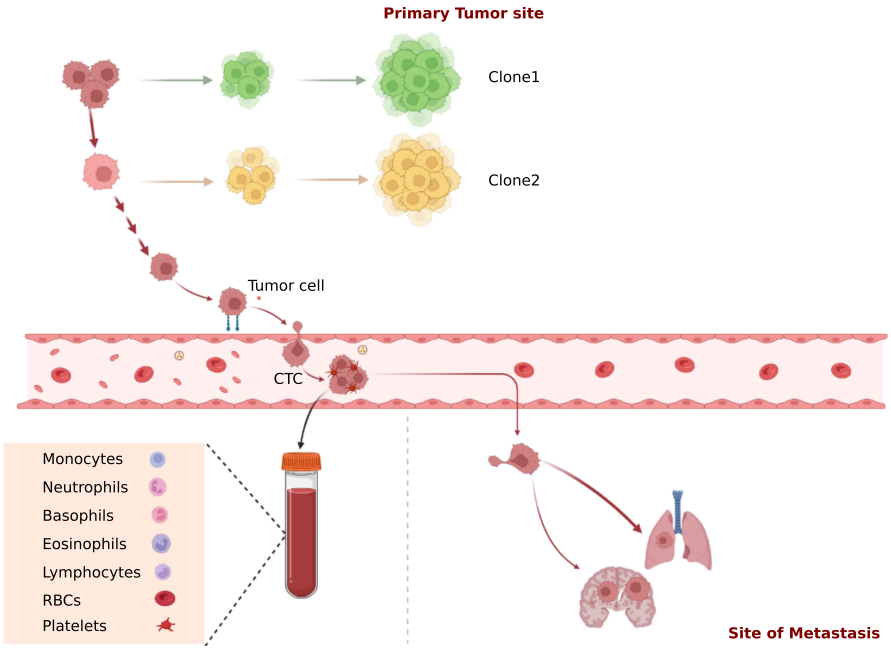

Figure 1.3: Metastasis: A multistep journey. Metastasis in cancer is a multistep process that involves several interdependent stages. These include the invasion of tumor cells at the primary site, entry into the circulation through intravasation, survival as CTCs while interacting with blood cells, exit from the circulation through extravasation, and establishment of a secondary site through attachment and colonization.

1.2.3.2 Isolating and analyzing circulating tumor cells

Recently, CTCs have gained increasing importance because of their multi-potential uses. Despite their long-known discovery and spates in clinical oncology, no method has been devised to isolate or enumerate CTCs efficiently. Primarily, their quantity in blood circulation is the biggest hurdle in isolating CTCs. Out of the several CTCs shed by the primary tumor, only about 0.1% survive in the circulation, and only about 0.01% are responsible for metastasis (Balic *et al.*, 2005). It has been reported that CTCs are not continuously shed in circulation. They are discontinuous and might not be present in homozygous conditions. Thus, while isolating CTCs, a single blood sample might fall insufficient or give inaccurate results (Nesteruk *et al.*, 2014). This is accompanied by a further reduction in their numbers when they get clogged in capillaries due to their large size. They can also form clusters while flowing; some may even adhere to the walls of

the capillaries, or some might be cloaked by the platelets. Further reduction in CTCs number occurs during batch processes followed for their enrichment. More straightforward methods involve size-based separation, collagen adhesion method, or density-based separation. Other sophisticated ones rely on epithelial markers, immunomagnetic techniques, microchips, and nanotech approaches (Alix-Panabières and Pantel, 2014).

1.2.3.3 Harnessing the unique potential of circulating tumor cells for advancing cancer treatment

Circulating tumor cells have emerged as promising biomarkers for cancer diagnosis, prognosis, and treatment response monitoring (Kowalik *et al.*, 2017). Over the past few years, they have been doled with much research, understanding that CTCs are shed from tumors and circulate in the bloodstream (Pantel and Speicher, 2016). Unlike other liquid biopsies products, such as circulating cell-free DNA or extracellular vesicles, CTCs are intact, viable cells that retain the genetic and phenotypic features of the primary tumor (Ma and Jeffrey, 2020). This feature makes CTCs more informative and reliable for assessing tumors' genomic and molecular heterogeneity and identifying potential therapeutic targets (Ma and Jeffrey, 2020; Ignatiadis *et al.*, 2021). Moreover, CTCs can be isolated from the peripheral blood using different techniques, including microfluidic platforms and immunomagnetic separation. Their analysis can provide real-time information on the disease status and response to therapy (Bankó *et al.*, 2019). These advantages of CTCs make them invaluable in enhancing cancer therapy and tailored healthcare.

1.2.3.4 Clinical implications of circulating tumor cells in cancer diagnosis and treatment

CTCs have significant clinical implications in cancer diagnosis and treatment. CTCs are shed from primary tumors and circulate in the bloodstream, and their detection and analysis can provide valuable information about cancer progression, therapy response, and prognosis (Potdar and Lotey, 2015). CTCs are a non-invasive biomarker (Bulfoni *et al.*, 2016) that can be used to diagnose cancer and monitor treatment efficacy in real-time, providing a more comprehensive assessment of disease status than traditional imaging and biopsy methods (Ma and Jeffrey, 2020; Ignatiadis *et al.*, 2021). Moreover, CTC

analysis can identify specific molecular markers and mutations in individual cancer cells, enabling the selection of targeted therapies and personalized treatment strategies. The enumeration of CTCs has been shown to correlate with disease stage and patient survival in several types of cancer, including breast, lung, prostate, and colorectal cancer (Yang *et al.*, 2016). Overall, CTC analysis can improve cancer diagnosis, treatment, and monitoring, leading to better patient outcomes (Ma and Jeffrey, 2020; Ignatiadis *et al.*, 2021; Zhang *et al.*, 2021; Yap *et al.*, 2014).

1.2.3.5 Challenges and future directions in the Study of circulating tumor cells

Despite the potential clinical significance of CTCs, several challenges must be addressed to realize their full potential in cancer diagnosis and treatment. One major challenge is the rarity of CTCs in the bloodstream, making their detection and isolation difficult (Yao *et al.*, 2014; Poonia *et al.*, 2022). Additionally, CTCs are a heterogeneous population, varying in size, morphology, and molecular characteristics, which can complicate their analysis (Millner *et al.*, 2013; Poonia *et al.*, 2022; Ilie *et al.*, 2014). Furthermore, the detection and analysis of CTCs require advanced and specialized technologies, which can be costly and not readily available in clinical settings (Vasseur *et al.*, 2021). Despite these challenges, several promising technologies and strategies are being developed to overcome these limitations and advance the study of CTCs. These include the development of microfluidic devices and assays for CTC isolation and analysis (Descamps *et al.*, 2022), as well as the integration of CTC analysis with other liquid biopsy components for a more comprehensive assessment of cancer status (Alimirzaie *et al.*, 2019). Additionally, new approaches for the analysis of CTCs, such as single-cell sequencing and functional assays, are being developed to provide more detailed molecular and phenotypic information about individual CTCs (Thiele *et al.*, 2019).

1.3 Cancer treatment approaches

Over time, significant progress has been made in the field of cancer treatments. Nowadays, a wide array of treatment options are available, which are tailored to specific factors such as the cancer type and stage, the patient's age, and overall health. Typi-

cally, a combination of various treatment methods is used to maximize the effectiveness of the therapy. The following are the most frequently utilized approaches for cancer treatment.

1.3.1 Surgery for cancer treatment

Surgical procedures are frequently the initial therapy option for cancer. The primary objective of surgery is to remove the cancerous tumor. In addition, surgery may involve the removal of nearby lymph nodes for cancer cell examination (Matsuda *et al.*, 2017). Surgery can be used as a standalone treatment or in combination with other therapies, such as chemotherapy or radiation therapy (Debela *et al.*, 2021). Surgery has evolved significantly with advancements in surgical techniques and technology. For instance, minimally invasive surgeries like laparoscopic and robotic surgery are becoming more prevalent. These procedures enable surgeons to do operations with smaller incisions, less blood loss, and quicker recuperation periods (Dare *et al.*, 2015; Gill *et al.*, 2002; Ohuchida and Hashizume, 2013).

1.3.2 Radiation therapy for cancer treatment

Radiation therapy employs X-rays or other types of radiation with high energies to destroy cancer cells and shrink tumors. Radiation therapy can be administered alone or in conjunction with other therapies. The type and duration of radiation therapy depend on the stage and type of cancer (Mehta *et al.*, 2010). Radiation therapy has also seen significant advancements in recent years. For example, Proton therapy is a recently developed radiation therapy technique that utilizes protons to accurately target cancer cells, reducing the risk of damage to nearby healthy tissue (Mutter *et al.*, 2021).

1.3.3 Chemotherapy for cancer treatment

Chemotherapy is a cancer treatment that involves the use of drugs to destroy cancer cells. These drugs can be given through an IV, taken orally, or applied topically to the skin (Al-Ostoot *et al.*, 2021; Feng and Chien, 2003). Chemotherapy may be used alone or in combination with other treatments. The specific type and duration of chemother-

apy depend on the stage and type of cancer. Chemotherapy has also undergone significant developments in recent years. Newer drugs that are more targeted and have fewer side effects than traditional chemotherapy drugs have been developed. For example, antibody-drug conjugates are a type of chemotherapy that targets specific proteins in cancer cells, reducing the risk of damage to healthy cells (Ponziani *et al.*, 2020). One notable advancement in cancer treatment is the use of immune checkpoint inhibitors, a type of immunotherapy. These inhibitors, including pembrolizumab and nivolumab, have demonstrated remarkable efficacy in treating melanoma, a type of skin cancer. By targeting proteins such as PD-1 (programmed cell death protein 1) or CTLA-4 (cytotoxic T-lymphocyte-associated protein 4), immune checkpoint inhibitors enhance the immune system's capacity to identify and eliminate cancer cells. This approach has shown significant success in improving patient outcomes in melanoma treatment (Larkin *et al.*, 2015; Wolchok *et al.*, 2017).

1.3.4 Immunotherapy for cancer treatment

Immunotherapy is a cancer treatment that utilizes the body's immune system to fight cancer. It involves using drugs that target particular proteins in cancer cells, which activate the immune system to recognize and attack cancerous cells. It can be utilized as a monotherapy or in conjunction with other modalities (Zhou, 2014). Immunotherapy has revolutionized cancer treatment in recent years. It has treated several types of cancer, including melanoma, lung, and bladder. Immunotherapy has also been shown to have fewer side effects than traditional cancer treatments (Liu and Zeng, 2012).

1.3.5 Targeted therapy for cancer treatment

Targeted therapy employs medications that target particular proteins or genes that facilitate the growth and spread of cancer cells. Targeted therapy may be administered by itself or in conjunction with other treatments. Targeted therapy has seen significant advancements in recent years, with several new drugs approved for treating different cancer types. For example, tyrosine kinase inhibitors are a targeted therapy that blocks the action of specific proteins that promote cancer growth (Zhou and Li, 2022).

1.3.6 Combination therapy for cancer treatment

Combination therapy involves using two or more types of cancer treatment to improve the effectiveness of the treatment. For example, combination therapy may include surgery and chemotherapy, radiation therapy and immunotherapy, or a combination of chemotherapy drugs. Combination therapy aims to target cancer cells from multiple angles, making it more difficult for cancer to develop resistance to treatment (Gilad *et al.*, 2021).

1.3.7 Hormone therapy

For hormone-sensitive malignancies such as breast and prostate cancer, an effective treatment is available that targets specific hormones. Hormone therapy works by blocking the production or activity of hormones that fuel the growth and spread of cancer cells. Hormone therapy may be administered orally, by injection, or by implant (Deli *et al.*, 2020). Currently, various types of hormone therapies are available for cancer treatment. For example, anti-estrogen treatment is used to treat estrogen receptor-positive breast cancer, while androgen deprivation therapy is employed to manage prostate cancer (Desai *et al.*, 2021).

1.3.8 Stem cell or bone marrow transplantation

In this treatment, healthy stem cells are transferred into a patient to replace defective or damaged bone marrow. This treatment is used to treat cancers that affect the bone marrow, such as leukemia and lymphoma, as well as other types of cancer that have spread to the bone marrow (Simpson and Dazzi, 2019). There are two distinct types of procedures in the field of stem cell transplantation: autologous and allogeneic transplants. An autologous transplant employs the stem cells derived from the patient undergoing the transplant, whereas an allogeneic transplant utilizes stem cells obtained from a donor (Simpson and Dazzi, 2019; López-Larrea *et al.*, 2012).

1.3.9 Personalized therapy and precision medicines

Personalized therapy and precision medicine are two related approaches to cancer treatment that involve tailoring treatment to the individual characteristics of each patient and their cancer. Personalized therapy considers factors such as the patient's genetic makeup, the type and stage of cancer, and overall health (Seymour *et al.*, 2017; Adams and Petersen, 2016). Precision medicine uses advanced technology such as genomics, proteomics, and other -omics fields to analyze the patient's cancer at the molecular level, identifying specific genetic mutations or other biomarkers that can be targeted by treatment (Hirsch *et al.*, 2017). This approach allows for a more targeted and effective treatment strategy with fewer side effects (Drake *et al.*, 2018; Cardona *et al.*, 2020). Precision medicine has been shown to improve patient outcomes and reduce healthcare costs compared to traditional cancer treatments (Schwaederle *et al.*, 2015).

1.3.9.1 Advantages and successes of precision medicine in cancer treatment and management

- **Personalized Treatment:** It enables medical practitioners to personalize treatment plans for patients based on their specific genetic, environmental, and lifestyle characteristics. By considering individual variability, clinicians can tailor interventions to optimize therapeutic outcomes and minimize adverse effects (Gambardella *et al.*, 2020).
- **Outcomes:** By targeting treatments to the specific molecular characteristics of a patient's cancer, precision medicine can potentially improve outcomes and reduce side effects (van de Sant *et al.*, 2019).
- **Early Detection:** Precision medicine identifies high-risk patients for specific cancers, allowing early detection and intervention (van de Sant *et al.*, 2019).
- **Better Drug Development:** Precision medicine allows researchers to identify new drug targets and biomarkers, accelerating the development of new cancer therapies (Gambardella *et al.*, 2020).
- **Cost-Effective:** By targeting treatments to patients who are most likely to benefit, precision medicine can reduce healthcare costs and improve resource allocation (van de Sant *et al.*, 2019).

1.3.9.2 Future directions and challenges in precision medicine for cancer research

Despite its successes, precision medicine still faces several challenges, including the need for more accurate and reliable biomarkers, the difficulty of integrating complex

genomic data into clinical decision-making, and the high cost of developing and implementing new targeted therapies (Naithani *et al.*, 2021). Future research should focus on addressing these challenges and improving the effectiveness and accessibility of precision medicine for all cancer patients (Yang *et al.*, 2019; Madhavan *et al.*, 2018).

1.4 Multiple myeloma: Diagnosis and treatment

Multiple myeloma is a type of cancer that develops from the plasma cells of the immune system. These cells are responsible for producing antibodies that fight against infections. The incidence of multiple myeloma varies by geographic location, with higher rates occurring in developed countries. The incidence of multiple myeloma is higher in males than in females (Cowan *et al.*, 2022). Multiple myeloma accounts for approximately 1% of all cancer cases and 10% of hematologic malignancies. According to the American Cancer Society's statistics, in 2023, around 35,730 new cases of multiple myeloma will be identified, with an estimated 12,590 deaths resulting from the disease (Cowan *et al.*, 2022).

1.4.1 Traditional and emerging diagnostic methods for multiple myeloma

Diagnosing multiple myeloma can be challenging as it often presents non-specific symptoms such as fatigue, bone pain, and recurrent infections. Therefore, early diagnosis is crucial for effectively treating and managing the disease. Traditional diagnostic methods for multiple myeloma include Imaging studies such as X-rays, computed tomography (CT) scans, magnetic resonance imaging (MRI) scans, bone marrow biopsy, urine tests, and blood tests. These methods have limitations in terms of sensitivity and specificity, and emerging diagnostic methods are being developed to improve early detection and accuracy (Cowan *et al.*, 2022; Michels and Petersen, 2017).

One promising emerging diagnostic method is liquid biopsies, which involve analyzing cell-free DNA, RNA, and proteins in the blood to identify cancer-specific biomarkers. This approach is less invasive and can be used to track the progression of disease and treatment response. Another emerging method is imaging techniques such as positron emission tomography (PET) scans, which can detect changes in bone metabolism

and assess the extent of the disease (Pappas *et al.*, 2022). Additionally, genetic testing and next-generation sequencing (NGS) can help identify genetic mutations and chromosomal abnormalities associated with multiple myeloma, providing personalized treatment options for patients.

1.4.2 Autologous stem cell transplantation as a therapeutic modality for patients with multiple myeloma

Autologous stem cell transplantation is a standard treatment option for eligible multiple myeloma patients. The procedure involves collecting and storing the patient's healthy blood-forming stem cells, which are then infused into the patient's bloodstream after high-dose chemotherapy to replace the damaged cells. This approach helps restore the bone marrow and immune system, which are often damaged by chemotherapy (Morè *et al.*, 2022). The efficacy of ASCT in enhancing the survival outcomes of patients diagnosed with multiple myeloma has been demonstrated in many studies, especially those who have not responded to initial treatment or have relapsed after prior therapy. However, it may not be suitable for all patients, and careful patient selection is necessary to minimize the risk of complications such as infections, bleeding, and organ damage (Morè *et al.*, 2022). In conclusion, multiple myeloma is a rare and complex disease that requires early and accurate diagnosis for effective treatment and management. Traditional diagnostic methods have limitations in sensitivity and specificity. Emerging diagnostic techniques such as liquid biopsies, PET scans, and genetic testing are being developed to improve early detection and accuracy. Autologous stem cell transplantation is an established therapeutic modality for eligible patients, which has been shown to improve survival outcomes. However, careful patient selection and monitoring are necessary to minimize the risk of complications (Mohty and Harousseau, 2014).

1.5 Single-cell RNA sequencing: Unlocking the secrets of cancer at the cellular level

Cancer is a heterogeneous and complex disease from genetic mutations, epigenetic alterations, and environmental factors. ScRNA-seq is a powerful technique that facilitates

the investigation of individual cells' gene expression patterns, providing unprecedented insight into the heterogeneity and complexity of cancer cells. This section provides an overview of scRNA-seq and its applications in cancer research, diagnosis, and treatment (Haque *et al.*, 2017; Lei *et al.*, 2021).

1.5.1 Overview of single-cell sequencing

Single-cell sequencing technologies have rapidly evolved in the past decade, providing researchers with the ability to study gene expression profiles in individual cells. scRNA-seq is a high-throughput technique that enables the analysis of thousands of cells in a single experiment (Jovic *et al.*, 2022). The workflow typically involves the isolation of individual cells, converting RNA to cDNA, library preparation, and sequencing on a high-throughput platform. The resulting data can be used to identify subpopulations of cells, study cell differentiation pathways, and identify novel cell types (Haque *et al.*, 2017; Moreno *et al.*, 2021). Various scRNA-seq technologies have been developed to analyze gene expression in individual cells. These include techniques that utilize droplets, microwells, and nanowells (Salomon *et al.*, 2019; Gole *et al.*, 2013; Goldstein *et al.*, 2017). These methods differ in efficiency, throughput, and cost. However, all share the same basic workflow of converting RNA to cDNA, followed by library preparation and sequencing (Moreno *et al.*, 2021).

1.5.2 Applications of single-cell RNA sequencing in cancer research

Diagnosis and Treatment scRNA-seq has numerous applications in cancer research, diagnosis, and treatment. One of the most significant applications is the identification of subpopulations of cancer cells within a tumor. scRNA-seq can identify some rare cell populations that may be missed by traditional bulk RNA sequencing (Fa *et al.*, 2021). The identification of these rare cell populations can provide insight into the heterogeneity of the tumor and may help to identify new targets for therapy (Watson *et al.*, 2018; Wu *et al.*, 2021). scRNA-seq can also be used to study the differentiation pathways of cancer cells (Das *et al.*, 2022). By identifying the gene expression profiles of cancer cells at different stages of differentiation, researchers can gain insight into the molecular mechanisms underlying cancer progression (Martin *et al.*, 2013). This information

can be used to develop novel biomarkers for diagnosis and prognosis along with new therapeutic targets (Yap *et al.*, 2014). In addition, scRNA-seq can be used to study the tumor microenvironment and plays a crucial role in cancer progression and response to therapy. scRNA-seq can identify the different cell types present in the tumor microenvironment and their gene expression profiles (Zhang *et al.*, 2021; Christensen *et al.*, 2022). This information can be used to develop novel immunotherapies that target specific cell types in the tumor microenvironment (Xiao and Yu, 2021). Overall, scRNA-seq is a powerful tool that can provide unprecedented insight into the heterogeneity and complexity of cancer cells. By identifying subpopulations of cancer cells, studying differentiation pathways, and characterizing the tumor microenvironment, scRNA-seq can provide new targets for therapy and enhance the overall well-being of patients (Lei *et al.*, 2021).

1.6 Machine learning: Revolutionizing cancer detection and treatment

Machine learning is an artificial intelligence technique in which algorithms are trained to assimilate information and make decisions or predictions based on the acquired knowledge (Huang *et al.*, 2020; Sharma and Prabha, 2021). In recent years, researchers have been exploring the potential of machine learning to improve cancer detection, diagnosis, and personalized treatment (Sharma *et al.*, 2018; Wang *et al.*, 2020*b*). Machine learning has several advantages in these areas, and several techniques are being used for personalized cancer therapy and management (Sudha *et al.*).

1.6.1 Advantages of machine learning in understanding cancer phenotypes

Machine learning has the ability to rapidly and precisely examine vast quantities of data, identify patterns that may be missed by human analysis, and make predictions about a patient's response to treatment. In cancer diagnosis, machine learning can analyze medical images, such as mammograms or MRI scans, to identify tumors or other abnormal findings (Bi *et al.*, 2019). For example, a recent study used a deep learning al-

gorithm to analyze mammograms and found that it outperformed human radiologists in detecting breast cancer (Cheng *et al.*, 2016; Kooi *et al.*, 2017). In cancer treatment, machine learning can help personalize treatment plans by predicting a patient's response to different therapies. This is important because not all patients respond the same way to a particular therapy. By analyzing the medical history of patients, genomic data, and other clinical information, machine learning algorithms have been utilized for feature selection, clustering, and dimensional reduction techniques, which aid in identifying relevant information from large datasets, improving the efficiency and accuracy of data analysis, and ultimately contributing to the development of personalized medicine, can identify the best treatment options for a particular patient. For example, a study using machine learning to predict the response of patients with metastatic colorectal cancer to different chemotherapy regimens found that the algorithm's predictions were more accurate than those of human oncologists (Huang *et al.*, 2020; Rafique *et al.*, 2021).

1.6.2 Unveiling complex cancer phenotypes through advanced learning techniques

1.6.2.1 Leveraging supervised learning

Sophisticated supervised learning algorithms, such as Support Vector Machines (SVM) and Random Forest, exhibit substantial promise in characterizing both known and unidentified cancer phenotypes. By utilizing labeled datasets, these algorithms reveal intricate patterns and characteristics inherent in recognized phenotypes. SVM and Random Forest efficiently categorize novel phenotypes by assessing similarities to established patterns, significantly advancing our comprehension of diverse cancer phenotypes.

1.6.2.2 Exploring via unsupervised learning

Additionally, unsupervised learning methodologies significantly contribute to unraveling complex cancer phenotypes. Techniques including clustering and dimensionality reduction provide insight into patterns and relationships without the reliance on labeled data. For instance, clustering methods like K-means or Hierarchical Clustering unveil intrinsic structures within cancer datasets. Dimensionality reduction techniques, such as

Principal Component Analysis (PCA) or t-distributed Stochastic Neighbor Embedding (t-SNE), facilitate the visualization of intricate molecular relationships among different cancer types. These unsupervised techniques provide a deeper understanding of latent and intricate connections, aiding in the characterization of diverse cancer phenotypes.

1.6.3 Machine learning techniques for personalized cancer therapy and management

Machine learning techniques have revolutionized cancer therapy and management by allowing for more personalized and effective treatments (Azuaje, 2019). These techniques use algorithms to analyze large datasets of patient information, including imaging, genomic, and clinical data, to identify patterns and predict outcomes. By leveraging machine learning algorithms, clinicians can make more informed decisions about treatment options, optimize therapy regimens, and elevate the overall health of patients. (Bottaci *et al.*, 1997; Rafique *et al.*, 2021).

1.6.3.1 Imaging analysis

Imaging analysis is critical to cancer diagnosis, treatment planning, and monitoring. Machine learning techniques have been used to analyze medical images such as CT scans, MRIs, and PET scans to identify tumors, measure their size and shape, and track their response to treatment. These techniques can also be used to identify biomarkers and genetic signatures associated with specific types of cancer, which can help clinicians tailor treatments to individual patients (Jasti *et al.*, 2022).

1.6.3.2 Genomic analysis

Genomic analysis involves analyzing a patient's DNA to identify mutations and other genetic abnormalities that may contribute to the development and progression of cancer. Machine learning techniques have been used to analyze large genomic datasets, allowing researchers to identify new biomarkers and potential drug targets (Vamathevan *et al.*, 2019; Zhu and Dupuy, 2022). By analyzing a patient's genomic profile, clinicians can personalize treatment plans, choosing the most likely effective therapies

and minimizing side effects (Rafique *et al.*, 2021; Lin *et al.*, 2021).

1.6.3.3 Electronic health records (EHRs)

EHRs are digital patient health information records, including medical history, medications, allergies, and test results (Ehrenstein *et al.*, 2019). Machine learning techniques can be used to analyze EHRs to identify patterns and trends in patient health, such as disease progression over time or certain treatments' effectiveness (Shinozaki, 2020; Turner, 2019). By leveraging machine learning algorithms, clinicians can make more informed decisions about treatment options and improve patient outcomes (Atasoy *et al.*, 2019).

1.6.3.4 Clinical decision support systems (CDSSs)

CDSSs are computer-based tools that help clinicians make informed decisions about patient care (Wasylewicz and Scheepers-Hoeks, 2019). CDSSs use machine learning techniques to analyze patient data, including treatment histories, test results, and imaging studies, to provide recommendations for diagnosis and treatment (Zikos and DeLellis, 2018; Peiffer-Smadja *et al.*, 2020). By incorporating CDSSs into clinical practice, clinicians can improve the accuracy of diagnoses and treatment plans, reduce errors, and elevate the overall health of patients (Sutton *et al.*, 2020).

1.7 Statistical methods for predicting cancer prognosis and treatment outcomes

Statistical methods play a crucial role in predicting cancer prognosis and treatment outcomes (Roychoudhury and Lahiri, 2018). By analyzing large data sets and identifying patterns, statisticians can help clinicians make informed treatment decisions, leading to better patient results.(Wiuf and Andersen, 2009). The use of survival analysis (Clark *et al.*, 2003), predictive modeling, risk stratification, meta-analysis, and personalized medicine can aid in the development of more effective and personalized treatment plans for cancer patients (Bodard *et al.*, 2023). The integration of statistical methods in can-

cer research has the potential to improve treatment outcomes and reduce the burden of cancer (Roychoudhury and Lahiri, 2018).

1.7.1 Comprehending survival analysis: The non-parametric estimation of Kaplan-Meier

The Kaplan-Meier estimator is a widely used non-parametric method for estimating survival probabilities in cancer patients by taking into account the time-to-event data of patients, including time to death or disease progression (Kaplan and Meier, 1958). It estimates the probability of survival at specific time points and provides valuable insights into cancer prognosis and treatment outcomes (Goel *et al.*, 2010). Kaplan-Meier analysis is a powerful tool for predicting cancer prognosis and treatment outcomes, as it accounts for the different lengths of follow-up and censored data, which is common in cancer research (Glinsky *et al.*, 2005). One of the key benefits of Kaplan-Meier analysis is that it enables the estimation of survival curves for various patient subgroups, enabling researchers to identify patient characteristics associated with better or worse outcomes (Garbe *et al.*, 1995). Additionally, it provides the median survival time, a useful metric for comparing treatments or patient subgroups (Goel *et al.*, 2010).

Furthermore, Kaplan-Meier analysis estimates the likelihood of remaining disease-free over time, which is crucial for predicting cancer recurrence and identifying patients at high risk of disease progression (Caplan *et al.*, 1994; Goswami *et al.*, 2019). By integrating Kaplan-Meier analysis into cancer research, clinicians and researchers can enhance their ability to make informed treatment decisions, leading to better results for patients. For instance, Kaplan-Meier analysis can be employed to estimate the probability of survival for patients with different stages of cancer, various treatment regimens, or distinct genetic mutations (Bollschweiler, 2003; Nagy *et al.*, 2021; Rosen *et al.*, 2020; Li *et al.*, 2023). This information serves as a guide for treatment decisions and contributes to optimizing patient outcomes. In summary, Kaplan-Meier analysis proves invaluable for predicting cancer prognosis and treatment outcomes, empowering clinicians and researchers to make more enlightened decisions that benefit patients (Lacny *et al.*, 2015; Morgera *et al.*, 2002; Deyo *et al.*, 2000).

1.7.2 The role of Bayesian network analysis in cancer prognosis and treatment outcomes

Another statistical method used in cancer prognosis and treatment prediction is the Bayesian network. Bayesian networks are probabilistic graphical models that represent the relationships between variables and can be used to make predictions based on available data (Heckerman, 1998). The Bayesian network model is based on Bayes' theorem, which uses prior knowledge and new evidence to update the probability of an event (Liu *et al.*, 2016). In the context of cancer research, Bayesian network analysis can be used to integrate clinical, genetic, and other data to predict the likelihood of disease progression, treatment response, and overall survival (Ni *et al.*, 2014; Agrahari *et al.*, 2018; Arora *et al.*, 2019; Ladyzynski *et al.*, 2022; Kraisangka and Druzdzel, 2018). One of the main advantages of Bayesian network analysis is that it can incorporate multiple data sources to make predictions. This includes clinical data such as age, tumor size, and tumor stage and genetic data such as gene expression profiles and somatic mutations. By integrating these data sources, Bayesian network analysis can provide more accurate predictions of patient outcomes than traditional statistical models that only consider one type of data (Ni *et al.*, 2014; Wang *et al.*, 2013). Another benefit of Bayesian network analysis is that it can be used to identify the most important variables for predicting patient outcomes (Genc and Dag, 2016). This can be helpful for guiding treatment decisions and identifying biomarkers that may be useful for predicting treatment response. Additionally, Bayesian network analysis can be used to identify patient subgroups that are at high risk of disease progression, or that may benefit from more aggressive treatment.

1.8 Scope of the thesis

The incidence of cancer has been increasing, and it has become a primary concern due to the lack of promising treatments for its management. The inter and intra-tumor heterogeneity and constant tumor evolution due to internal and external environmental exposure make it challenging to develop effective treatment options. Early detection and appropriate management of cancer can significantly increase the chances of successful treatment. Machine learning has dramatically contributed to diagnosis, treatment, and

management in the last two decades. The recent advancements in liquid biopsy and single-cell transcriptomics have opened new avenues to track cancer heterogeneity and identify the clinical factors responsible for treatment failure or relapse. This thesis aims to use machine learning techniques to tackle these significant challenges. Our approach is to develop statistical and computational-based algorithms to identify and characterize circulating tumor cells, which are live cells responsible for metastasis and the primary reason for cancer-related deaths. Moreover, we focus on identifying other clinical factors that make cancer treatment worse or fail. The research outlined below aims to significantly contribute to cancer treatment management by creating valuable resources for early detection and personalized treatment plans that can improve patient outcomes.

1.8.1 Unsupervised identification of circulating tumor cells from a pool of white blood cells

Traditional methods for identifying CTCs using single-cell RNA sequencing face significant limitations due to the molecular diversity of cancer cells, the low concentration of CTCs in blood, and the down-regulation of epithelial markers (Ju *et al.*, 2022; O'Flaherty *et al.*, 2017). To address these challenges, we develop a marker-agnostic approach, which employs deep dictionary clustering to detect CTCs in an unsupervised manner from a pool of white blood cells. This innovative approach transforms normalized and log-transformed expression vectors linked to CTCs into a pathway enrichment score vector, which is then used to cluster CTCs and white blood cells into distinct clusters. The method is robust, neutralizes batch effects, and offers a functional/mechanistic perspective on unraveling cellular heterogeneity, making it a valuable tool for future CTC characterization studies.

1.8.2 In silico characterization of circulating tumor cells

We utilized various computational and statistical tools to analyze CTC transcriptomes and annotate clusters obtained through unsupervised clustering in comparison to those of white blood cells. One method employed was the inference of copy number variation (CNV), which enabled cluster annotation by identifying the malignant origin of CTCs and pinpointing the precise position of chromosomal aberrations. Additionally, we uti-

lized single markers and sets of markers to check the expression of CTCs and immune cells. However, due to the epithelial-mesenchymal transition (EMT) and other physical and chemical stressors, CTCs may lose some of their epithelial properties, making univariate differential expression analysis less useful. To address this, we employed Stouffer's method, which allowed for the collective measurement of several canonical markers identifying the malignant, immune, or epithelial origin. We found that gene set-based approaches can enhance the characterization of cell clusters based on single-marker and inferred CNV approaches.

1.8.3 Unraveling clinical factors influencing relapse in myeloma patients undergoing ASCT

Effective cancer management is crucial for improving patient outcomes in the field of oncology. However, one of the significant challenges in cancer treatment is the risk of relapse, which can be caused by various clinical factors. This study aims to identify the critical clinical factors responsible for early relapse in patients with multiple myeloma who underwent ASCT . The study examined clinical data from 253 patients who underwent treatment at the Department of Medical Oncology in the All India Institute of Medical Sciences (AIIMS), India. Advanced statistical techniques, including decision tree analysis and Bayesian networks, were used to identify primary disease status, diabetes, and Glomerular Filtration Rate (GFR) as critical factors affecting early relapse. These findings can provide valuable insights to clinicians in identifying high-risk patients and developing effective treatment strategies to enhance overall patient outcomes.

CHAPTER 2

Unsupervised clustering of circulating tumor cells from a large pool of blood cells

2.1 Introduction

Cancer is a major contributor to global mortality rates and poses a significant challenge to increasing life expectancy worldwide (Sung *et al.*, 2021). In 2019, according to the report of the World Health Organization (WHO), cancer was the primary cause of death for individuals of age under 70 years in 183 countries and ranked as the third or fourth leading cause in 23 countries (Mathers, 2020; Sung *et al.*, 2021). Metastasis is the fundamental reason for 90% of deaths because of cancer. Metastasis occurs when cancer cells segregate from the primary tumor and enter the bloodstream and populate at some different organ and where these cells colonize distant organs and spread the malignancy (Bittner *et al.*, 2020; Krebs *et al.*, 2014; Siegel *et al.*, 2015). Cancer cells secrete chemokines to attract immune cells, promoting tumor proliferation and intravasation. After cancer cells enter the bloodstream, they face a huge nuisance, but some manage to leave the vasculature and reach a secondary site (Shenoy and Lu, 2016; Follain *et al.*, 2018). CTCs have gained significant attention in recent years for their role in tumor metastasis. These cells can be found in 40% to 80% of patients with metastatic breast cancer and offer clinically relevant information for cancer diagnosis and treatment (Hong *et al.*, 2016; Kwa and Esteva, 2018). A high abundance of CTCs in peripheral blood is associated with a poor disease prognosis (Bork *et al.*, 2015; Cristofanilli *et al.*, 2004; Danila *et al.*, 2007; Rack *et al.*, 2014; Giuliano *et al.*, 2011; Tsai *et al.*, 2016). Metastasis is facilitated by epithelial-to-mesenchymal transition (EMT), which enables tumor epithelial cells to acquire mesenchymal-like features and easily enter the bloodstream (Bulfoni *et al.*, 2016). Single-cell RNA sequencing plays a crucial role in the unsupervised characterization of CTC transcriptomes (Ranjan *et al.*, 2021; Kiselev *et al.*, 2019; Chen *et al.*, 2020; Guo *et al.*, 2015; Kiselev *et al.*, 2017; Macosko *et al.*,

2015; Butler *et al.*, 2018; Wolf *et al.*, 2018). Up until now, the majority of single-cell RNA sequencing studies on CTCs have relied on marker-based methodologies for segregating CTC subpopulations, while marker-agnostic techniques for CTC annotations are limited and cannot confirm that the cells are cancerous. This is mainly due to the following reasons: 1. high levels of molecular heterogeneity among malignant cells (Li *et al.*, 2017; Tirosh *et al.*, 2016). 2. The low concentration of CTCs (0–10 CTCs per mL of blood) in peripheral blood (Alix-Panabières and Pantel, 2013). 3. The down-regulation of epithelial markers through the EMT process (Mikolajczyk *et al.*, 2011; Iyer *et al.*, 2020). 5. Batch effects across various scRNA-seq studies (Kiselev *et al.*, 2019; Büttner *et al.*, 2019). We introduce the unCTC R package to address the challenges of identifying circulating tumor cells using scRNA-seq technologies. This software package contains a variety of computational and statistical tools that enable an independent analysis of CTC transcriptomes in comparison to those of WBCs. The package contains both standard and novel modules for clustering, inference of copy number variation (CNV), and marker-dependent recognition of CTC and immune cell groups. For clustering, we present the DDLK technique, which efficiently separates CTC and WBC populations using pathway scores at the single-cell level and deep dictionary learning. Our technique includes multidimensional analyses of single-cell expression profiles to assign a phenotypic identity to the captured cells.

2.2 Methodology

2.2.1 Description of datasets

In our analysis, we employed seven separate scRNA-seq count data from CTCs and WBCs. We downloaded these datasets from various sources (Velten *et al.*, 2017; Ting *et al.*, 2014; Yu *et al.*, 2014; Sarioglu *et al.*, 2015; Jordan *et al.*, 2016; Aceto *et al.*, 2014; Zheng *et al.*, 2017). Out of these, six datasets provided 141 single CTCs, while two other studies generated 1037 WBCs. It's noteworthy that one dataset (accession number: GSE67939) encompassed both blood and CTC transcriptomes (Sarioglu *et al.*, 2015), while the other dataset (accession number: GSE74639) included six single primary tumor cells and ten single CTCs (Zheng *et al.*, 2017). The CTC data we utilized involved three types of cancer, specifically lung, pancreatic, and breast cancer, as shown

in **Table 2.1**. We utilized these datasets to validate the capability of unCTC for integrative analysis and clustering.

Table 2.1: A detailed explanation of the datasets.

Dataset	Tissue	WBCs/ PBMCs	CTCs	CTC Clusters	Primary Tumor Cells
Aceto et al. (GSE51827) (Aceto *et al.*, 2014)	Breast	0	15	14	0
Jordan et al. (GSE75367) (Jordan *et al.*, 2016)	Breast	0	74	0	0
Sarioglu et al. (GSE67939) (Sarioglu *et al.*, 2015)	Breast	2	15	0	0
Ting et al. (GSE60407) (Ting *et al.*, 2014)	Pancreatic	0	7	0	0
Velten et al. (GSE75478) (Velten *et al.*, 2017)	Bone marrow (HSCs and progenitors)	1035	0	0	0
Yu et al. (GSE55807) (Yu *et al.*, 2014)	Breast	0	6	0	0
Zheng et al. (GSE74639) (Zheng *et al.*, 2017)	Lung	0	10	0	6

2.2.2 Data integration, filtration, and normalization

In this study, a data integration approach was used to combine seven different datasets based on common genes. The resulting integrated dataset was subjected to filtration to remove genes and cells with low expression values. To improve the reliability of the results, we excluded cells that expressed fewer than 1500 genes, as determined by a read count greater than zero. Likewise, we removed genes that exhibited non-zero expression in fewer than five cells. We utilized the Linnorm normalization approach to account for differences in gene expression across different batches and individual cells, applying its default parameters for batch correction and single-cell normalization (Yip *et al.*, 2017). Following the normalization step, we applied a logarithmic transformation to the expression values, adding 1 as a pseudo-count.

2.2.3 Expression matrix to pathway enrichment score matrix

To compute scores for gene-set enrichment, we used the R software package Gene Set Variation Analysis (GSVA) (Hänzelmann *et al.*, 2013). GSVA requires two inputs: the log-transformed and normalized expression matrix and the gene sets. We obtained the gene sets from the C2 collection in the MSigDB database (Subramanian *et al.*, 2005), which consists of over 6000 gene sets that were curated from literature sources. To eliminate biases in the enrichment score, we filtered out genes from the C2 gene set collection that were absent from the expression matrix. We specified a minimum gene set the size of 10, a maximum size of 500, and set max.diff as FALSE. To speed up the computing procedure, we calculated enrichment scores concurrently using four distinct threads (*parallel.sz=4*). This was feasible due to the fact that computations for each set of genes are not interdependent.

2.2.4 DDLK clustering

The GSVA enrichment score matrix was utilized as the input data for unsupervised clustering. We used *k-means*-compatible deep dictionary learning (DDL) and the elbow technique to determine the optimal number of clusters for clustering. The DDLK technique requires three inputs: the number of clusters, pathway enrichment scores, and the path to Python 3 to conduct computations and generate clusters.

The equation presented below is the standard mathematical expression for *k-means* clustering:

$$\sum_{i=1}^{k}\sum_{j=1}^{n} h_{ij}\|z_j - \mu_i\|^2, \tag{2.1}$$

where $h_{ij} = 1$ if x_j is in cluster i, and $h_{ij} = 0$ otherwise.

Here, z_j is the jth sample, and μ_i is the centroid of the ith cluster.

Matrix factorization provides an alternative mathematical representation of *k-means* clustering (Bauckhage, 2015).

$$\|Z - ZH^T(HH^T)^{-1}H\|_F^2 \tag{2.2}$$

In this work, we prefer the expression of *k-means* using the formula 2.2, which involves the use of Z as the data matrix, where the columns are produced by vertically assembling z_j's and matrix H that contains the binary indicator variables h_{ij}.

Given that DDL is not a widely used framework, we will provide a brief overview. Dictionary learning (Tošić and Frossard, 2011) involves learning a base matrix (D) which can be used to generate or synthesize data (X) from its corresponding coefficients (Z).

$$X = DZ \tag{2.3}$$

While the notion of dictionary learning is a recent development, the underlying issue has previously been referred to as matrix factorization. In equation (2.3), we can observe that the data matrix X is factorized into D (base matrix) and Z. The following expression demonstrates the underlying solution to the matrix factorization/dictionary learning problem –

$$\min_{D,Z} \|X - DZ\|_F^2 \tag{2.4}$$

where $||\cdot||_F^2$ denotes the squared Frobenius norm, which is calculated by summing up the squares of every element within the matrix.

In DDL, the approach of learning multiple layers of the dictionary is used instead of learning a single layer. This can be represented as

$$X = D_3\phi(D_2\phi(D_1Z)) \tag{2.5}$$

In DDL, multiple layers of dictionaries are learned, represented by D_1, D_2, D_3, and so on, with an activation function ϕ between two layers. In this work, three layers are used as an example, but the number of layers can be more. This process can be visually represented through a flowchart. **Figure 2.1** illustrates the matrix factorization of the input pathway matrix up to three layers of Deep Dictionary Learning (DDL), providing

a visual depiction of the successive transformations through each layer in the flowchart.

Unsupervised formulation's solution can be expressed as follows –

$$\min_{D_1,D_2,D_3,Z} \|X - D_1\phi(D_2\phi(D_3Z))\|_F^2 \tag{2.6}$$

In our study, the Rectified Linear Unit (ReLU) activation function was chosen for two key reasons. Firstly, it can be easily integrated into the optimization constraint. Additionally, ReLU has superior function approximation capabilities. Hence, we will express our basic DDL framework (with ReLU) as follows,

$$\min_{D_1,D_2,D_3,Z} \|X - D_1D_2D_3Z\|_F^2 \; s.t. \underbrace{D_2D_3Z \geq 0, D_3Z \geq 0, Z \geq 0}_{\text{ReLU activation}} \tag{2.7}$$

We suggest that DDL formulation 2.7 should include the *k-means* cost 2.2. Our approach involves utilizing the DDL-generated features as inputs for clustering; rather than addressing the problem separately, we optimize the following cost function in a unified manner –

$$\begin{aligned} \min_{D_1,D_2,D_3,Z,H} \underbrace{\|X - D_1D_2D_3Z\|_F^2}_{\text{DDL loss}} + \mu \underbrace{\|Z - ZH^\top(HH^\top)^{-1}H\|_F^2}_{\textit{k-means} \text{ loss}} \\ s.t. \underbrace{D_2D_3Z \geq 0, D_3Z \geq 0, Z \geq 0}_{\text{ReLU activation}} \end{aligned} \tag{2.8}$$

In equation 2.8, the parameter μ is used to balance the weight between the dictionary learning and *k-means* clustering loss. In order to assign equal importance to both loss functions, we set μ=1. Therefore, the final formulation is presented as follows:

$$\min_{D_1,D_2,D_3,Z,H} \underbrace{\|X - D_1D_2D_3Z\|_F^2}_{\text{DDL loss}} + \underbrace{\|Z - ZH^\top(HH^\top)^{-1}H\|_F^2}_{\textit{k-means loss}} \quad s.t. \underbrace{D_2D_3Z \geq 0, D_3Z \geq 0, Z \geq 0}_{\text{ReLU activation}} \tag{2.9}$$

To solve equation 2.9, we employ an alternating minimization approach. Initially, we disregard the non-negativity constraints stated in equation 2.9, but we will address them later. The updates for various components of the variable are outlined below,

$$D_1 \leftarrow \min_{D_1} \|X - D_1D_2D_3Z\|_F^2 \tag{2.10}$$

$$D_1^k = XZ_1^\dagger, \ where \ Z_1 = D_2^{k-1}D_3Z^{k-1} \tag{2.11}$$

Here, $Z1^\dagger$, the cross superscript ($\dagger$) represents the Pseudoinverse.

$$D_2 \leftarrow \min_{D_2} \|X - D_1D_2D_3Z\|_F^2 \tag{2.12}$$

$$D_2^k = (D_1^k)^\dagger XZ_2, \ where \ Z_2 = D_3^{k-1}Z^{k-1} \tag{2.13}$$

$$D_3 \leftarrow \min_{D_3} \|X - D_1D_2D_3Z\|_F^2 \tag{2.14}$$

$$D_3^k = (D_1^kD_2^k)^\dagger X(Z_{k-1})^\dagger \tag{2.15}$$

$$Z \leftarrow \min_{Z} \|X - D_1D_2D_3Z\|_F^2 + \|Z - ZH^\top(HH^\top)^{-1}H\|_F^2 \tag{2.16}$$

To resolve Z, it is necessary to compute the gradient of the equation presented in 2.16 and set it equal to zero. The process of deriving this solution is explained subsequently.

$$\nabla(\|X - D_1D_2D_3Z\|_F^2 + \|Z - ZH^\top(HH^\top)^{-1}H\|_F^2) = 0 \tag{2.17}$$

$$\implies (D_1D_2D_3)^\top X - (D_1D_2D_3)^\top(D_1D_2D_3)^\top Z - Z(I - H^\top(HH^\top)^{-1}H) = 0 \tag{2.18}$$

$$\implies (D_1D_2D_3)^\top X = (D_1D_2D_3)^\top (D_1D_2D_3)^\top Z + Z(I - H^\top (HH^\top)^{-1}H) \quad (2.19)$$

Equation 2.19 indicates that Z satisfies the Sylvester equation in the form of AZ + ZB = C, where A = $D_1D_2D_3^\top (D_1D_2D_3^\top)$, B = $(I - H^\top (HH^\top)^{-1}H)$ and C = $(D_1D_2D_3)^\top X$. We can solve the Sylvester equation in 2.19 to get the update for Z.

$$\min_H \|Z - ZH^\top (HH^\top)^{-1}H\|_F^2 \quad (2.20)$$

The last stage involves resolving equation 2.20 in order to modify H. To obtain the revised H, we can use *k-means* clustering on the updated Z.

Thus far, the ReLU non-negativity constraints have not been incorporated into the derivation. In practice, enforcing these constraints would require employing forward-backward type splitting algorithms, which are iterative and could significantly increase the algorithm's runtime. Instead, we satisfy the constraints by setting any negative values in Z, Z_1, and Z_2 zero during each iteration.

The algorithm is presented concisely below. Once the algorithm converges, the clusters can be identified from H. However, since equation 2.8 is a non-convex function, convergence cannot be guaranteed. Thus, we terminate the iterations once H remains relatively unchanged in subsequent iterations.

Initialize: $D_1^0, D_2^0, D_3^0, Z_0, H_0$
repeat
  Update D_1^k, D_2^k, D_3^k using (2.11), (2.13), (2.15)
  Update Z_k by solving Sylvester's equation in 2.19
  Update H_k by applying *k-means* clustering on the updated Z
until *convergence*;

Algorithm 1: DDLK

2.2.5 Differentially expressed genes identification

To identify the genes that were differentially expressed among the clusters generated from DDLK clustering, we utilized the voom approach (Law *et al.*, 2014) in conjunction with the Limma package (Ritchie *et al.*, 2015). The voom approach was specifically chosen for its capability in addressing heteroscedasticity commonly found in RNA-seq data, and its effectiveness in modeling variance. Moreover, its suitability for studies

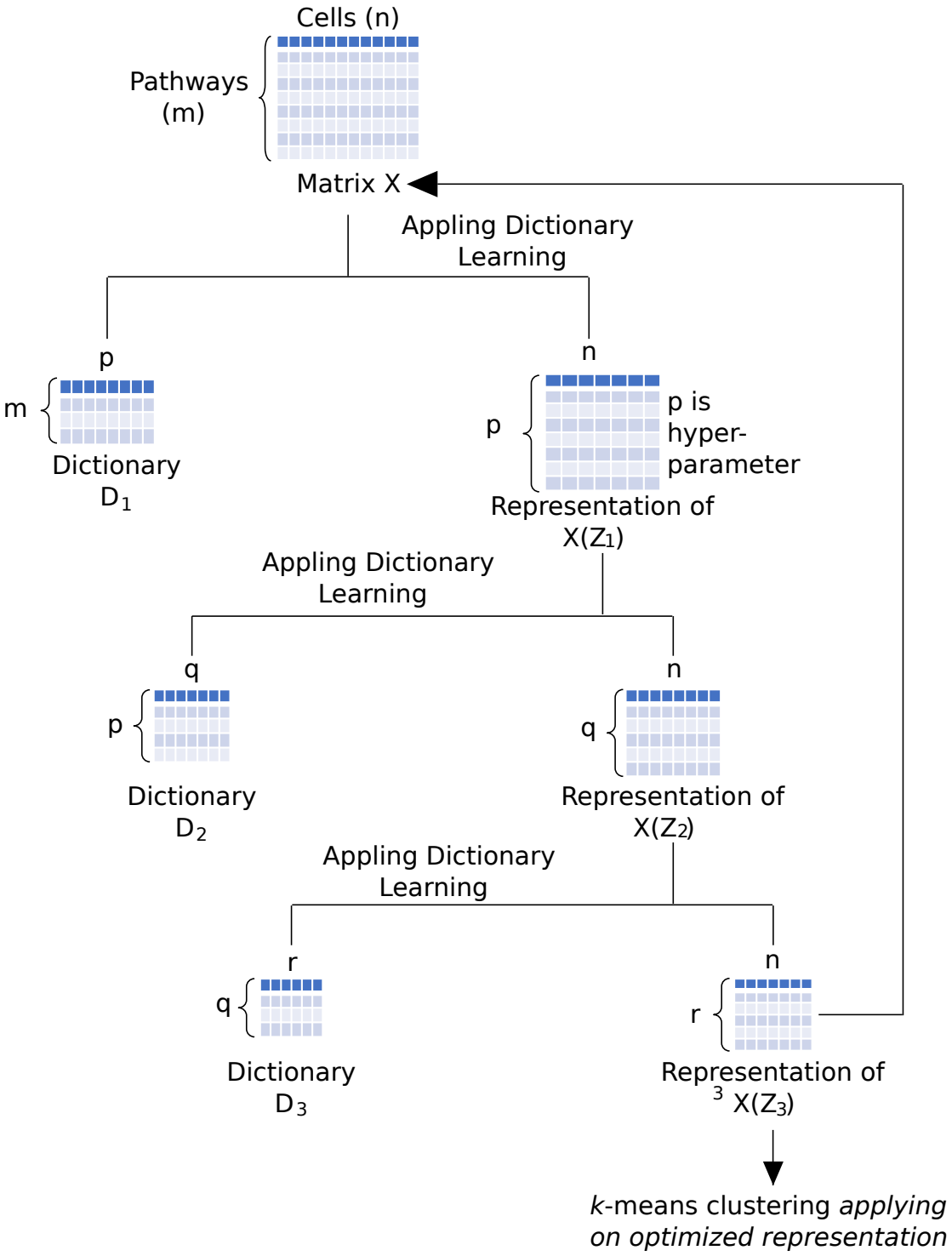

Figure 2.1: Dictionary learning, a technique analogous to matrix factorization. Matrix factorization is applied to a pathway matrix where rows denote pathways and columns represent individual cells, denoted as the input matrix 'X.' This process involves the factorization of the original matrix X into two components: a dictionary, referred to as D_1, and coefficients, labeled Z_1. Successive dictionary learning iteration is applied to Z_1 until complete optimization is achieved. Rectified Linear Unit (ReLU) activation functions are employed at each layer. The resultant features derived from Deep Dictionary Learning (DDL) are inputs for subsequent clustering analysis. Rather than an incremental approach, a joint optimization method is implemented using a specific cost function. It's crucial to note the non-convex nature of the problem, which may limit reaching a global minimum. Complete convergence is identified when significant changes in convergence are absent.

with smaller sample sizes and its robustness in handling RNA-seq data were crucial factors influencing our selection. However, it's important to note that voom may be sensitive to outliers and could be computationally intensive. In comparison to other normalization and transformation methods such as TMM (trimmed mean of M-values) (Robinson and Oshlack, 2010) or RPKM (Reads Per Kilobase Million), voom's strength in stabilizing variance and addressing the unique characteristics of RNA-seq data made it the preferred choice for our study.

Initially, we constructed a DEGlist object using the normalized expression matrix. This object was then processed through the *calcNormFactors()* function of an R package, edgeR (Robinson *et al.*, 2010; Computing), with *method* = *none* and *normalization factor* = *1*, followed by a Voom transformation. In order to identify the most significantly up-regulated genes in every cluster, we sorted the results based on log fold change and utilized a threshold of 0.05 for the adjusted $p-value$. Further functional analysis of the up-regulated genes unique to each cluster was conducted using Ingenuity Pathway Analysis (IPA) (Krämer *et al.*, 2014).

2.2.6 Upregulated differential pathways

To assess the distinctive pathways between the clusters obtained from DDLK clustering, we employed the Limma package (Ritchie *et al.*, 2015). For differential pathway analysis, we used the moderated t-statistic. Pathways exhibiting a positive log fold change and an adjusted $P - value$ less than 0.05 were distinctly enriched in a particular cell group.

2.2.7 Assessment of lineage markers via combinatorial analysis

For combinatorial analysis of gene sets, we used Stouffer's score (Stouffer *et al.*, 1949). Stouffer's method involves putting together Z-scores from a number of features to get a single score that shows how much of a certain trait or characteristic is present (Gupta *et al.*, 2020). In this study, we examined the number of lineage-indicating marker genes using Stouffer's approach. These contain genes for immune and epithelial cell types. Individual markers may not demonstrate noteworthy statistical significance concerning differential expression, which makes this method particularly suitable for CTC analysis.

2.2.8 Copy number variation analysis

To determine the copy number variation (CNV) profile of individual cancer cells, we utilized the inferCNV R package (Tickle *et al.*, 2019). To obtain copy number aberrations in circulating tumor cells, we used the count profile of peripheral blood mononuclear cells (PBMCs) from healthy individuals as a reference. The inferCNV plot showed that the test group and the control group had significant copy number variations (CNVs) in their scRNA-Seq profiles. We used GRCh37's cytoband information to find out where a specific gain or loss happened on the p- or q-arm of a chromosome. In reference cells, the minimum average number of reads count per gene was set to 1 (Barrios and Prieto, 2017).

2.2.9 Seurat, fastMNN, and Harmony's execution details

We evaluated the performance of unCTC in comparison to three established integrative analysis methods, namely Seurat, fastMNN, and Harmony. Seurat version 4.1.1 was used for two different approaches to integrating single-cell studies (Butler *et al.*, 2018). The first approach, called Vanilla Seurat, involved combining scRNA-seq datasets using common genes to create the input gene expression matrix. We used *NormalizeData()*, *FindVariableFeatures()*, and *ScaleData()* functions for normalization, identification of highly variable genes, and scaling the data, respectively, with default parameters. Next, we applied *FindNeighbors()* and *FindClusters()* functions for clustering cells. For TPM normalized input data, we skipped the normalization step since the matrices were already length normalized and log-transformed the data without the normalization step. In the second approach, we used Seurat's integrative analysis pipeline called Integrative Seurat, which employs Reciprocal Principal Component Analysis (RPCA) and Canonical Correlation Analysis (CCA). We used *NormalizeData()* to normalize count data and log-transformed TPM data (skipping the normalization step). To integrate CTC and WBC datasets from various studies, *SelectIntegrationFeatures()* was used to identify informative features (genes), followed by the computation of anchors using the *FindIntegrationAnchors()* function. We used CCA and RPCA as reduction methods to find integration anchors with *k.anchor=3, dims=1:5*, and *k.score=5* parameters to avoid errors due to a small number of samples in either dataset. The *IntegrateData()* function

was then used to integrate these anchors with *k.weight* and *dims* parameters adjusted based on the smallest dataset's sample size. Finally, clustering was done using *FindNeighbors()* and *FindClusters()* functions with their default parameter settings.

The second approach that we employed is the Fast Mutual Nearest Neighbors Correction (fastMNN), which effectively addresses batch effects in single-cell expression data by utilizing a rapid implementation of the Mutual Nearest Neighbors (MNN) methodology (Haghverdi *et al.*, 2018). For fastMNN, we combined all WBCs and CTCs into a single scRNA-seq matrix based on common genes. We used the Seurat preprocessing workflow to filter, normalize, and scale the data. In the case of transcripts per million (TPM) data as input, we log-transformed the combined matrix by adding one pseudo count and omitted the normalization step. The *RunFastMNN()* function was used on the output from the previous step. We split the combined Seurat objects into a list of multiple Seurat objects based on the data sources. Clustering was performed using the *FindNeighbors()* and *FindClusters()* functions. Finally, we evaluated and visualized the batch-corrected outputs in the Uniform Manifold Approximation and Projection (UMAP) space.

The third approach we used, Harmony, was applied to align cells from distinct batches using an iterative clustering technique (Korsunsky *et al.*, 2019). To combine all WBCs and CTCs into a single scRNA-seq matrix based on common genes, we followed the Seurat preprocessing steps: filtering, normalization, and scaling of the data. Finally, we applied the *RunHarmony()* function on the output from the previous step with the parameter *group.by.vars* = *data source*. Additionally, we utilized the *FindNeighbours()* and *FindClusters()* functions for cell clustering. We evaluated the batch-corrected outputs by visualizing them in the UMAP space.

2.3 Results

2.3.1 Outline of the unCTC workflow

The dynamic nature of the CTC phenotype makes it difficult to identify and characterize CTCs using scRNA-seq profiles. To address this challenge, the unCTC workflow offers several methods for unbiased detection and description of single circulating tumor cell

transcriptomes. To achieve this goal, it is essential to cluster the scRNA-seq profiles accurately, and a reliable method is proposed for clustering single-cell transcriptomes in a metaspace. This method includes pathways and their enrichment scores, which are calculated based on the gene expression readouts of single cells. Single-cell expression data is often sparse (Kiselev *et al.*, 2019; Tian *et al.*, 2019), but pathway scores computed on gene sets help in the detection of cellular subtypes (Chawla *et al.*, 2021; Li *et al.*, 2017). To cluster CTCs in an unsupervised manner, each expression vector is first normalized and log-transformed. Then, pathway enrichment scores are computed using the gene set variation analysis (GSVA) (Hänzelmann *et al.*, 2013) to convert the expression vector into pathway enrichment scores. This transformation successfully eliminates batch effects (Kim *et al.*, 2018) and uncovers functional heterogeneity within cells (Ding *et al.*, 2019; Wang *et al.*, 2020*a*; Ramirez *et al.*, 2020; Ding *et al.*, 2020). DDLK is an example of semi-supervised clustering that integrates the *k-means* clustering cost into the Deep Dictionary Learning (DDL) framework. DDLK can overcome the limitations of dictionary learning and deep learning techniques, such as superficial learning and data dependency, by mapping single-cell gene expression data onto various well-known biological pathways, resulting in reliable cellular clusters (Tariyal *et al.*, 2016). While DDLK can identify comparable cellular subpopulations from scRNA-seq data with CTC expression profiles, characterizing these subpopulations may still present difficulties. To overcome this problem, we incorporated the inferCNV feature into the unCTC R package (Computing, n.d.). InferCNV is a proven technique that uses scRNA-seq data to estimate copy number variation (CNV) (Couturier *et al.*, 2020), and its integration helps to annotate the identified cell groups. In single-cell analysis, InferCNV is beneficial in detecting approximate CNV regions, particularly in cancer cells that experience significant chromosomal alterations (Durante *et al.*, 2020; Zhou *et al.*, 2020). These alterations can be precisely located at the level of chromosomal arms by inferCNV, which uses cytoband data based on GRCh37. This helps in identifying mutated genes and understanding the malignant source of CTCs (Barrios and Prieto, 2017). Because of the epithelial-mesenchymal transition (EMT) and other physical and chemical stressors, CTCs tend to lose some of their epithelial properties as they move to distant organs. This can make univariate differential expression analysis less useful. To address this constraint, the unCTC approach employs Stouffer's method, which allows for the collective measurement of a number of canonical markers that identify

the malignant, immune, or epithelial origin (Stouffer *et al.*, 1949). In our experience, gene set-based approaches can enhance the characterization of cell clusters based on single-marker and inferred CNV approaches. **Figure 2.2** outlines the complete unCTC workflow.

2.3.2 Effective separation of CTCs and WBCs utilizing DDLK clustering

The primary purpose of the unCTC technology is to distinguish between the CTC and WBC populations derived from the microfluidic enrichment of CTCs in blood and other body fluids. This objective differs from detecting CTC clusters or deciphering intrinsic heterogeneity within single-cell transcriptomes; it typically involves the unsupervised grouping of expression vectors. To accomplish this, gene expression vectors were projected onto a meta-space of well-defined biomolecular pathways using Gene Set Variation Analysis (GSVA). When supplied with selected pathways, GSVA converts the expression matrix into the pathway enrichment score matrix (Hänzelmann *et al.*, 2013), which makes it suitable for data integration tasks (Jin *et al.*, 2014). The resulting pathway enrichment score vectors are used for clustering the scRNA-seq transcriptome profiles. Current clustering techniques which are based on deep learning utilize stacked autoencoders (Xie *et al.*, 2016; Yang *et al.*, 2017; Fard *et al.*, 2020) or their convolutional counterparts (Yang *et al.*, 2017; Guo *et al.*, 2015). However, these methods have been shown to suffer from overfitting and degradation of results (Liang and Liu, 2015). Deep Dictionary Learning (DDL) is a more appropriate alternative when there is a limited amount of data available (Mahdizadehaghdam *et al.*, 2019; Tariyal *et al.*, 2016; Tang *et al.*, 2020; Fu *et al.*, 2019). We used the DDLK clustering technique, which combines the cost of *k-means* clustering into the DDL framework, to facilitate the clustering of data of different sizes. Most single-cell studies combine scRNA-seq data from multiple biological replicates, which leads to significant batch effects (Sinha *et al.*, 2019). Single-cell data also exhibits cell-to-cell technical variability due to a low level of starting RNA. Therefore, it is necessary to ensure the single-cell pipeline is resistant to these variance factors. To evaluate this, we generated a complex multi-study dataset (Ting *et al.*, 2014; Sarioglu *et al.*, 2015; Zheng *et al.*, 2017; Velten *et al.*, 2017; Yu *et al.*, 2014; Jordan *et al.*, 2016; Aceto *et al.*, 2014) consisting of 141 CTCs from pancreatic,

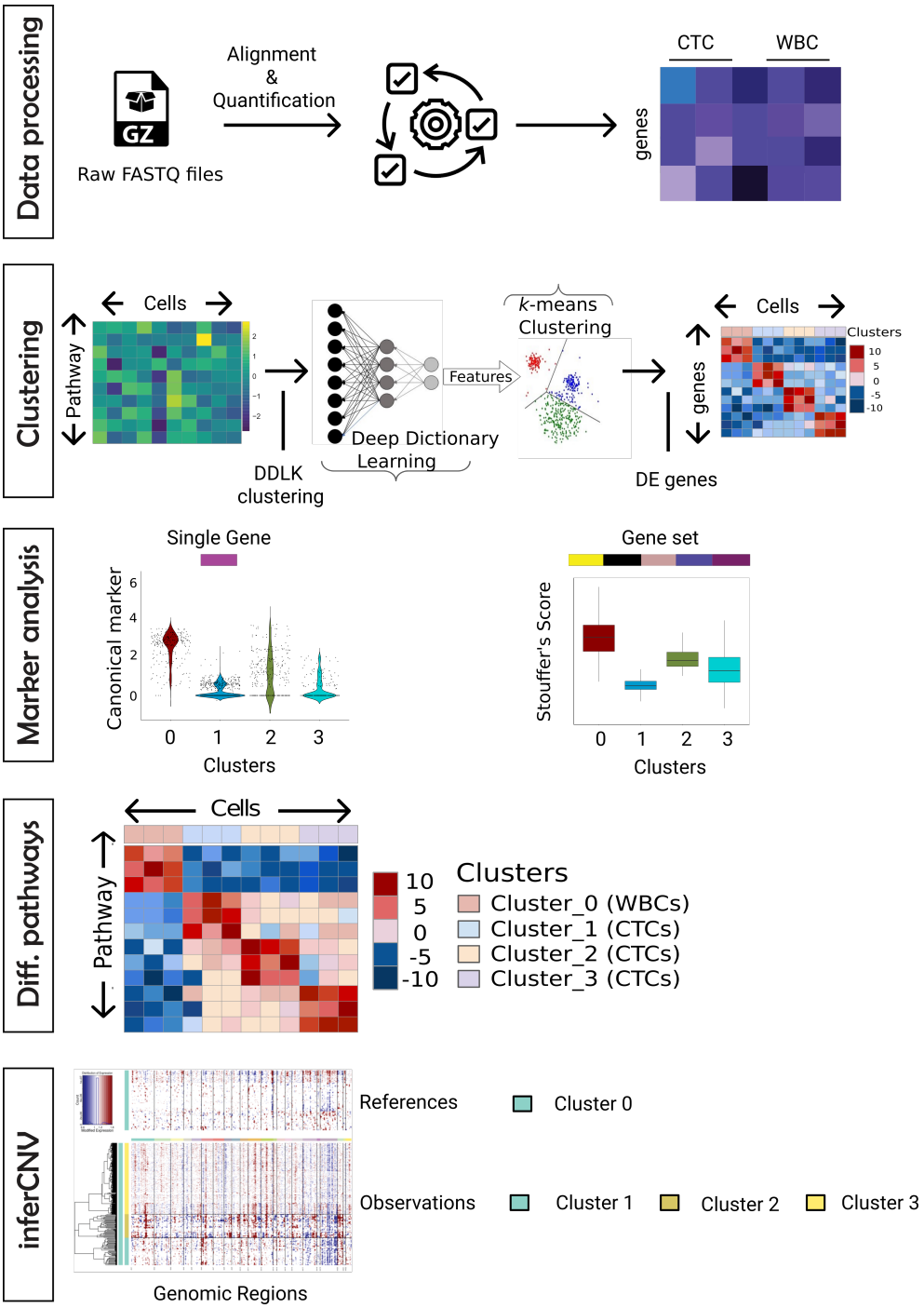

Figure 2.2: unCTC workflow: A comprehensive computational framework designed for the marker-free characterization of circulating tumor cells. The initial step is to generate an expression matrix by processing the raw FASTQ files. A clustering method called DDLK is used to cluster single CTC transcriptomes, using pathway enrichment scores to achieve robust clustering. Subsequently, cluster-level differential expression analysis provides insights into CTC and white blood cell subtypes. Immune and epithelial marker expression levels guide the estimation of general cell type identities. Further analysis involves well-known gene sets and pathways, providing a comprehensive understanding. In addition, functional properties are inferred through the analysis of pathway-specific gene expression, and expression-based pseudo-CNV inference aids in unbiased cancerous cell detection.

lung and breast cancers, as well as 1037 WBCs, to compare the performance of unCTC against fastMNN, Seurat, and Harmony **(Table 2.2)**. The study utilized two Seurat software variants: Vanilla Seurat and Integrative Seurat. Vanilla Seurat uses a consolidated matrix created by combining individual scRNA-seq datasets based on common genes. Integrative Seurat, on the other hand, utilizes Canonical Correlation Analysis (CCA) or Reciprocal Principal Component Analysis (RPCA) to merge multiple studies. CTCs and WBCs were visually separated by unCTC and Vanilla Seurat, whereas Harmony and fastMNN failed to distinguish between the two cell types **(Figure 2.3)**. The performance of various clustering approaches in terms of CTC/WBC separation is depicted in **Figure 2.3 A, B, D, E, G, H, J, and K**. unCTC and Vanilla Seurat could distinguish CTCs as belonging to distinct clusters **(Figures 2.3C and 2.3L)**. In contrast, clusters returned by fastMNN and Harmony had both sorts of cells **(Figure 2.3 F, I)**.

The Integrative Seurat approach, which employs Canonical Correlation Analysis and Reciprocal Principal Component Analysis (RPCA), has also been unable to distinguish between CTCs and WBCs successfully and instead shows clusters with a mixture of both cell types. **Figure 2.4** presents a UMAP-based visualization, clustering, and cluster purity assessment using both CCA **(Figure 2.4A-C)** and RPCA **(Figure 2.4D-F)** to showcase the performance of Integrative Seurat utilizing both methods. To objectively evaluate the quality of clusters obtained through other methods employed in this study, we utilized three commonly used metrics - cluster purity, Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). Our findings revealed that the unCTC method surpassed current methods in these metrics, demonstrating its superior performance **(Figure 2.4G)**.

2.3.3 Identifying cell lineages of DDLK clusters using markers

Although clustering is widely used for unsupervised multivariate analysis of single cells, determining the cell lineage through identifying marker genes requires evaluating the expression of those genes. Normally, scientists rely on univariate statistics and a small set of standard markers to determine the lineage of CTCs and WBCs. However, the behavior of these markers in CTCs can be uncertain and hard to anticipate. To avoid relying too heavily on specific genes, examining the expression of multiple markers unique to a particular cell lineage is beneficial. In this study, We employed Stouffer's

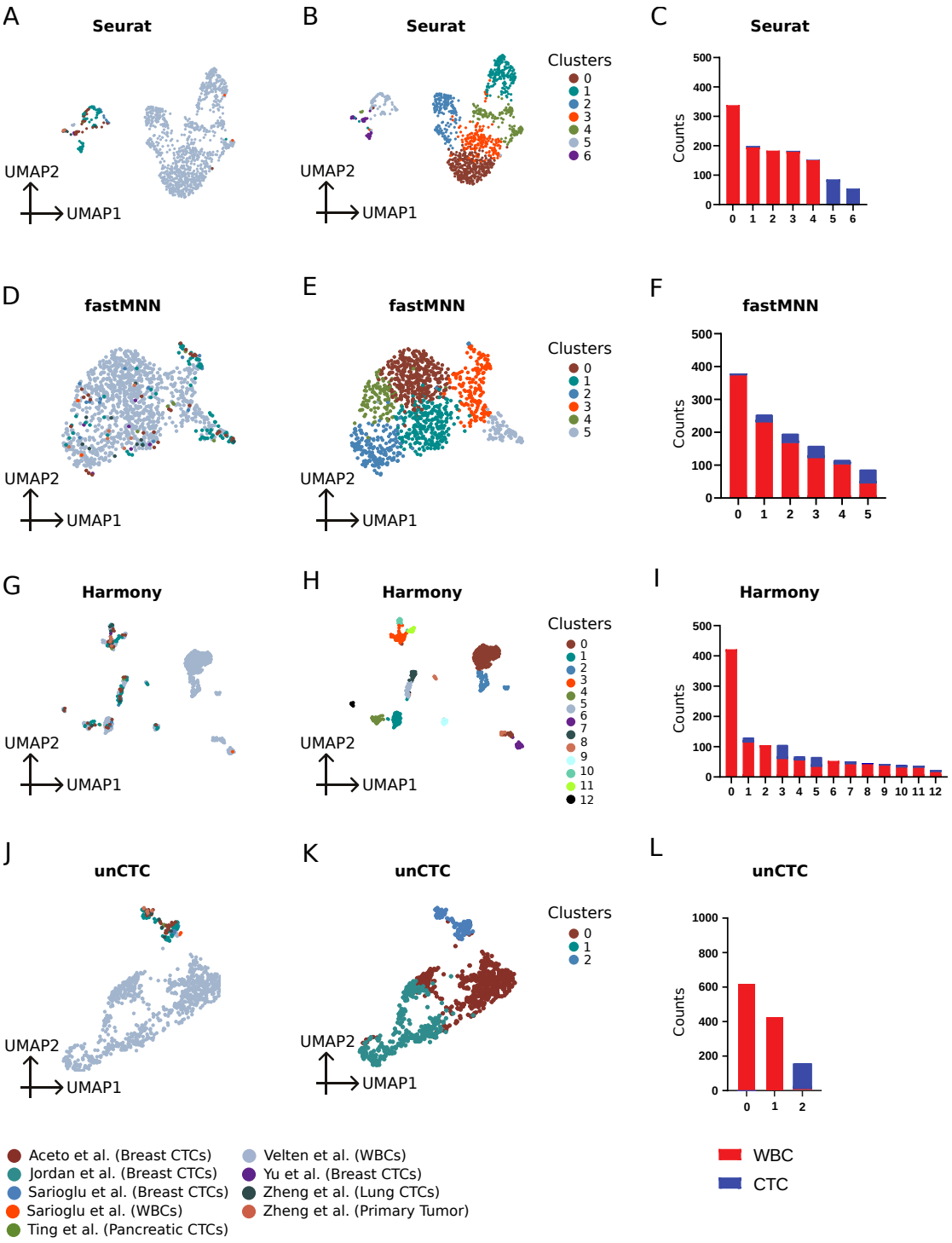

Figure 2.3: The unCTC approach facilitates the combined analysis of CTCs and WBCs. To evaluate the effectiveness of unCTC, a dataset containing CTCs and WBCs from multiple studies was utilized. The performance of unCTC was then compared against three of the most commonly used scRNA-seq analysis pipelines: Vanilla Seurat, fastMNN, and Harmony. The results of this analysis are presented in Figure 2.2, comprising panels A-C, which include A) a two-dimensional UMAP visualization generated using Vanilla Seurat, representing data from seven studies with a color code legend provided below the figure, B) 2D UMAP-based clusters of all seven study datasets, and C) an assessment of cluster purity for CTCs and WBCs. Panels D-F and G-I exhibit similar visualizations for fastMNN and Harmony, respectively, while panels J-L present the corresponding figures for unCTC.

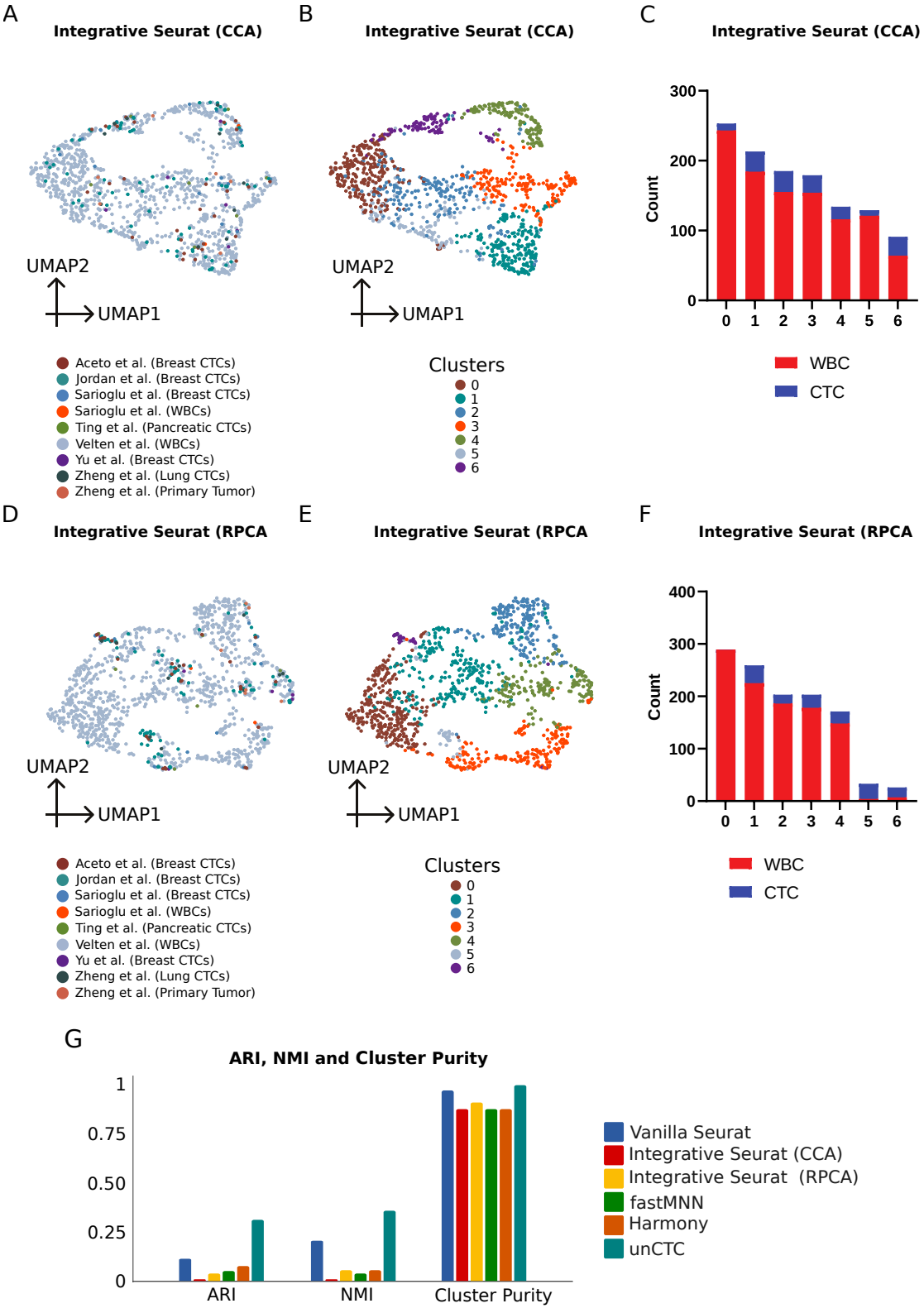

Figure 2.4: Seurat's integrative analysis. The integration of multiple datasets from different studies was analyzed using Seurat's canonical correlation analysis (CCA) and reciprocal principal component analysis (RPCA) methods, as depicted in Figure 3 (A-C) and (D-F), respectively. (G) Provides a summary of the performance of various methods employed in this study, based on comparing clusters with WBC/CTC annotations.

method to merge the expression data of several markers linked to the cell lineages of significance explained in Supplemental Table S4 of Poonia et al. study (Poonia *et al.*, 2022). In **Figures 2.5A-C**, we can observe the enrichment scores for specific genes associated with immune and epithelial cells across distinct clusters. Remarkably, among the three clusters identified utilizing unCTC, Cluster 2 displayed the highest enrichment of epithelial markers. This outcome is in line with the designations that were derived from previous research.

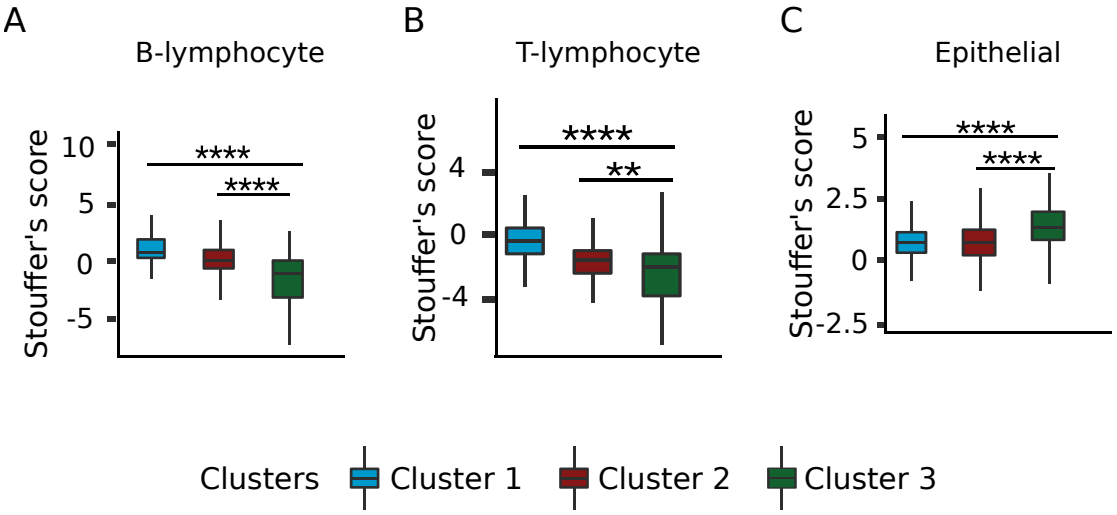

Figure 2.5: Purity of the clusters. (A-C) Boxplots representing the distribution of Stouffer's scores for recognized markers of epithelial cells, T cells, and B cells for each cluster identified using unCTC.

2.3.4 Analysis of HNSCC scRNA-seq data from 18 patients using unCTC

The popularity of single-cell RNA-seq technologies for characterizing the tumor microenvironment is on the rise, and differentiating between malignant and nonmalignant stromal cell types remains a recurrent obstacle in the analysis of scRNA-seq data. Typically, CNV inference based on gene expression addresses this challenge, but CNVs are not guaranteed to be available or detectable at the expression level. To show how well unCTC works in these situations, the unCTC pipeline was used to examine the scRNA-seq profiles of 18 patients with primary head and neck squamous cell carcinoma (HNSCC). The information about these profiles was acquired from a study published by Puram and his team in 2017 (Puram *et al.*, 2017). The unCTC pipeline was

applied to single-cell RNA sequencing profiles obtained from 18 patients (Puram *et al.*, 2017). A total of 324 cells with unknown annotations were removed for the analysis, leaving 5578 cells, of which 2215 were identified as malignant and 3363 as non-malignant by the original authors. Applying unCTC-based analysis on the scRNA-seq data resulted in the clear distinction of malignant and non-malignant categories, identifying four distinct clusters **Figure 2.6**. The distribution of the number of malignant and non-malignant cells across the four clusters obtained from unCTC is depicted in **Figure 2.6C**. The clustering performance of unCTC was compared to Vanilla Seurat's, which showed that unCTC-based segregation of malignant vs. non-malignant subtypes yielded greater values of Normalized Mutual Information (NMI) and Adjusted Rand Index **(Figure 2.6D)**. Further analysis of Stouffer's scores of combined expression of cell type-specific markers revealed that epithelial markers were enriched in clusters 0 and 2, confirming the original malignant annotations of cells in these clusters **(Figure 2.6E)**. Notably, the unCTC method effectively distinguished between the cancerous and non-cancerous groups without requiring manual intervention **(Figure 2.6)**.

2.4 Discussion

Circulating tumor cells can provide valuable insight into the tumors they originated from and cancer progression. However, many current CTC enrichment approaches are limited because they can only identify CTCs that show typical epithelial markers. The analysis of molecular profiles of cell populations enriched for CTCs can be accomplished through single-cell expression studies. However, there currently needs to be a comprehensive computational tool to identify unknown or heterogeneous phenotypes using scRNA-seq data, including both CTCs and WBCs. unCTC addresses this issue by offering various semi-supervised and supervised methods to analyze transcriptomes of CTCs and WBCs. We presented a novel clustering approach called DDLK, which utilizes pathway enrichment scores to create reliable groupings of single cells, even when datasets are derived from different studies. This is especially beneficial since most single-cell studies include multiple replicates. It's important to note that DDLK is designed to discover broad groups within scRNA-seq data and is not intended to help identify diverse subpopulations. For that, our previously published articles about the dropClust software package (Sinha *et al.*, 2019). By using unCTC, we were able to

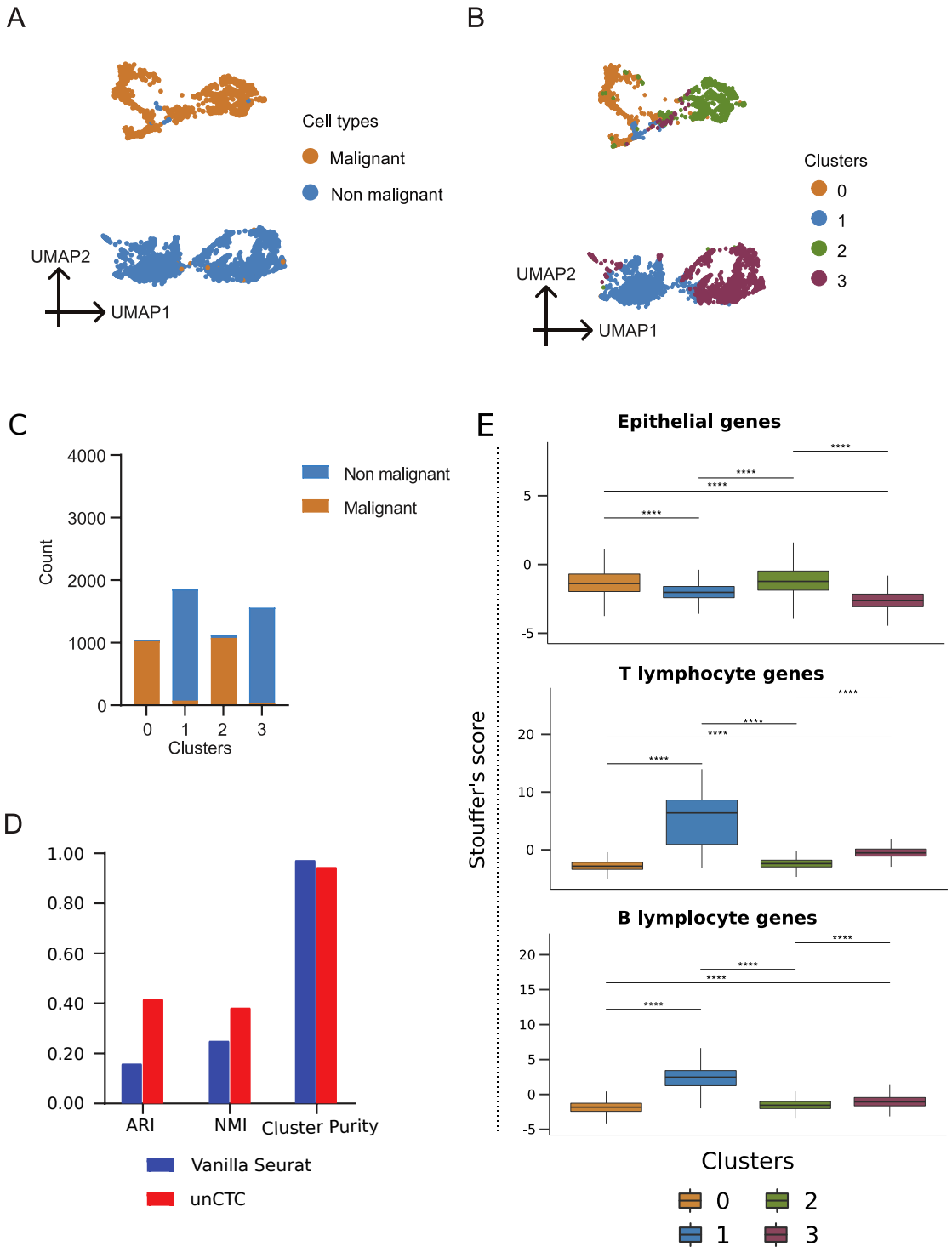

Figure 2.6: Clustering and Characterization of HNSCC Dataset using unCTC. (A) 2D UMAP visualization of the HNSCC expression dataset. (B) Clusters obtained through unCTC. (C) Bar plot illustrating the distribution of malignant and non-malignant cells in unCTC-derived clusters. (D) Bar plot showing cluster purity, ARI, and NMI for unCTC and Vanilla Seurat. (E) Box plots demonstrating the distribution of Stouffer's scores based on known markers of B cells, T cells, and epithelial cells for cells within each cluster identified by unCTC.

identify CTCs with unknown phenotypes. The study compared unCTC to four existing integrative single-cell analytic pipelines: Vanilla Seurat, Integrative Seurat (with both CCA and RPCA versions), fastMNN, and Harmony. Vanilla Seurat avoided merging both cell types efficiently **(Figure 2.3)**. In contrast, the integrated Seurat variants (CCA and RPCA) could not differentiate between CTC and WBC subpopulations **(Figure 2.4)**. Three frequently used metrics, namely cluster purity, Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI), were evaluated to objectively assess cluster qualities obtained using different methods. unCTC outscored the current methods for these metrics, indicating its superior performance **(Figure 2.4G)**. CTCs have become increasingly important in cancer diagnosis, prognosis, and treatment monitoring. These cells provide valuable insights into the diversity of tumors and the processes involved in their spread throughout the body. As the cost of single-cell expression profiling continues to drop, new methods like unCTC are likely to play a vital role in advancing our understanding of CTC biology. By constructing a cancer-specific molecular atlas of CTCs, researchers will be able to gain deeper insights into the mechanisms that govern CTC formation, dissemination, and survival.

DDLK utilizes deep dictionary learning with a k-means clustering cost for the analysis of large single-cell RNA sequencing (scRNA-seq) datasets. However, applying K-means clustering to extensive gene expression datasets introduces potential challenges, including scalability issues, computational efficiency concerns, and sensitivity to high dimensionality. To proactively overcome these challenges, we have implemented a thoughtful preprocessing step. Specifically, we transform the gene expression dataset into a Pathway score matrix before applying K-means clustering, effectively reducing its size and enhancing computational efficiency. This intentional preprocessing approach is strategically designed to optimize the performance of K-means within the unique context of our research. Recognizing the dynamic nature of clustering methodologies, we express our intent to explore alternative approaches in future studies. Contemplating hierarchical clustering and DBSCAN as viable options, these methods offer distinctive perspectives and potential solutions to enhance the clustering analysis in our research further. As the technology continues to improve and evolve, we can expect to see even more progress in the analysis of CTCs and their significance in cancer diagnosis, prediction, and management. Overall, using unCTC in CTC analysis is a promising development with the chance to improve our knowledge of cancer biology

and ultimately elevate patient outcomes.

2.5 Software availability

The most recent iteration of the analysis pipeline and code base for unCTC can be acquired via the website `https://github.com/SaritaPoonia/unCTC`. This resource provides researchers with access to the latest advancements and updates for the unCTC methodology.

CHAPTER 3

Employing unCTC for marker-free characterization of circulating tumor cells

3.1 Introduction

Liquid biopsy is a diagnostic technique that has gained popularity in recent years due to its non-invasive nature, and it has emerged as a promising tool for detecting cancer biomarkers. CTCs are one of the most promising biomarkers for liquid biopsy (Jin *et al.*, 2019*a*), providing valuable information about tumor characteristics and disease progression. Capturing CTCs is critical for monitoring cancer metastasis and response to treatment. Analyzing CTCs can lead to earlier detection, more accurate disease monitoring, and tailored therapies, ultimately resulting in better patient outcomes (He *et al.*, 2017; Maheswaran and Haber, 2010; Lin *et al.*, 2020). As a result, there is a rapidly growing interest in developing technologies and methods for capturing and analyzing CTCs. In recent times, various principles have been employed for CTC capture platforms, such as antibody-based capture (Riethdorf *et al.*, 2007; Nagrath *et al.*, 2007; Stott *et al.*, 2010), immune cell depletion (Ozkumur *et al.*, 2013), size exclusion (Xu *et al.*, 2015), and dielectrophoresis (Chiu *et al.*, 2016). Of these, CellSearch® is the first CTC capture technology that has been authorized by the FDA. It employs antibodies that target the EPCAM (Epithelial cell adhesion molecule) antigen in order to collect CTCs successfully (Iyer *et al.*, 2020; Habli *et al.*, 2020; Ignatiadis and Reinholz, 2011). The use of marker-based enrichment strategies is not ideal for identifying heterogeneous subpopulations of CTCs (Miller *et al.*, 2010; Farace *et al.*, 2011; Wang *et al.*, 2016). Previous studies have introduced alternative methods for capturing CTCs that utilize the physical properties of cancer cells (Gabriel *et al.*, 2016; Ferreira *et al.*, 2016; Cheng *et al.*, 2019), but the use of PTPRC negative selection as an alternative method for capturing CTCs has limitations (Gabriel *et al.*, 2016; Ferreira *et al.*, 2016). Negative selection involves removing cells expressing a specific marker, in this case, PTPRC, from the sample. This approach may increase the purity of CTCs, but immune cells that

lack PTPRC may still be captured along with CTCs, which can complicate downstream analyses and affect the accuracy of CTC enumeration. It is crucial to employ analytical methods that are independent of epithelial markers and allow for leukocyte contamination detection. This will facilitate the isolation of highly pure CTC populations, which can be further characterized using mRNA sequencing. In order to overcome this issue, we utilized a marker-free approach that involved the integration of the ClearCell® FX and Polaris™ systems. This allowed us to capture CTCs based on their size exclude immune cells, and subsequently perform single-cell sequencing to obtain gene expression profiles of putative CTCs (Warkiani *et al.*, 2014; Ramalingam *et al.*, 2016). We conducted an analysis of complete transcriptomes for a subset of CTCs derived from six patients with breast cancer of three distinct molecular subtypes ($ER^-/PR^-/HER2^-$, $ER^+/PR^+/HER2^-$, and $ER^-/PR^-/HER2^+$). We employed the unCTC analysis pipeline to perform this analysis and compared our results to those obtained from independent single-cell RNA sequencing of breast CTCs and leukocytes. Our findings verified the efficacy of the ClearCell-Polaris microfluidic approach for capturing CTCs without the use of markers. Furthermore, our study demonstrated the value of in silico characterization of CTCs in enhancing the process of marker-free CTC capture and characterization.

3.2 Description of datasets

We employed three scRNA-seq datasets in our study. Two of these datasets are publicly available and labeled as the Ebright et al. dataset (Ebright *et al.*, 2020) and the Ding et al. dataset (Ding *et al.*, 2020). The third dataset, called the Poonia et al. dataset, was generated in-house and consisted of 81 potential CTCs obtained from six breast cancer patients of three subtypes ($ER^+/PR^+/HER2^-$, $ER^-/PR^-/HER2^-$, and $ER^-/PR^-/HER2^+$) using the ClearCell® FX and Polaris™ workflow for CTC enrichment (Poonia *et al.*, 2022). As a control, we used the Ebright et al. dataset, which contained 824 ER^+/PR^+ single CTCs that were directly isolated from whole blood specimens of cancer patients using the CTC-iChip microfluidic system (Ebright *et al.*, 2020). The third dataset, Ding et al. dataset, contained a total of 752 white blood cell (WBC) expression profiles that were processed in two different runs (Ding *et al.*, 2020). We compared the gene expression levels of Poonia et al. CTCs and Ebright et al. CTCs, and found that Poonia et al. CTCs generally had higher gene expression levels, as illus-

trated in **Figure 3.1**. A description of the datasets used in this study is shown in **Table 3.1**.

Table 3.1: Description of datasets.

Dataset	#WBCs/PBMCs	#CTCs
Ebright et al. (GSE144494) (Ebright *et al.*, 2020)	0	824 breast CTCs
Poonia et al. (GSE186288) (Poonia *et al.*, 2022)		81 CTCs breast cancer
Ding et al. WBC1 (Ding *et al.*, 2019)	376	0
Ding et al. WBC2 (Ding *et al.*, 2019)	376	0

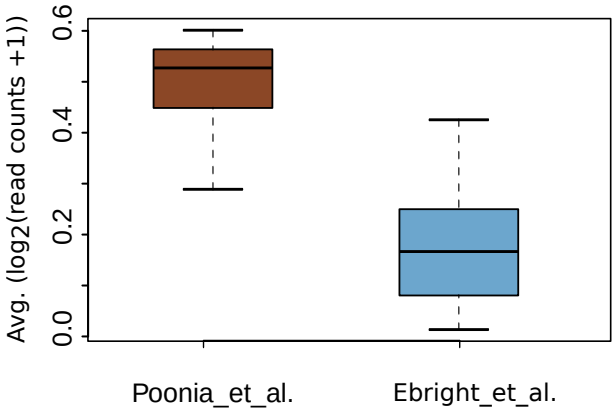

Figure 3.1: Exploring gene-specific read counts. The average gene-specific read counts across CTCs from both Poonia et al. and Ebright et al. datasets.

3.2.1 Workflow of marker-free capture of CTCs

3.2.1.1 Sample collection

A total of 81 CTCs were gathered from the blood samples of six breast cancer patients who had different molecular subtypes. Out of the six breast cancer patients with different molecular subtypes, the ones with TNBC contributed 11 CTCs, whereas those with $ER^+/PR^+/HER2^-$ subtype yielded 57 CTCs and patients with the $ER^-/PR^-/HER2^+$ subtype gave 13 CTCs (**Table 3.2**). The blood samples were collected by the National Cancer Center Singapore, and all the participants provided their informed consent as per the guidelines of the Institutional Review Board (IRB) (CIRB no. 2014/119/B). The clinical sample collection procedures were endorsed by the SingHealth Centralized Institutional Review Board. According to the most recent guidelines established by the College of American Pathologists and American Society of Clinical Oncology, the testing of progesterone receptor (PR), estrogen receptor (ER), and human epidermal

growth factor receptor 2 (HER2) status via immunohistochemical (IHC) methods was performed.

Table 3.2: ClearCell® FX and Polaris™ workflow-enabled CTCs: A detailed depiction.

Patient	#CTCs	Hormone Receptor Status
P1	11	ER-/PR-/HER2-
P3	13	ER+/PR+/HER2-
P4	12	ER+/PR+/HER2-
P5	32	ER+/PR+/HER2-
P7	5	ER-/PR-/HER2+
P9	8	ER-/PR-/HER2+
Total	81	

3.2.1.2 CTC enrichment

To perform the enrichment of circulating tumor cells, 9 milliliters of blood samples were obtained through the use of K3 EDTA blood collection tubes (Greiner Bio-One, 455036). To process each batch, a volume of 6-8.5 mL of whole blood was utilized. The first step involved eliminating red blood cells by adding a red blood cell (RBC) lysis buffer (G-Biosciences®, St. Louis, MO, USA), and the mixture was incubated at room temperature for 10 minutes. The centrifugation process was used to separate the lysed RBCs from the supernatant. Before CTC enrichment on the ClearCell FX system by Biolidics Limited, the nucleated cell pellet was resuspended in ClearCell resuspension buffer, following the instructions provided by the manufacturer (Lee *et al.*, 2018).

3.2.1.3 Immunofluorescence suspension staining

The CTC-rich blood samples were concentrated to a volume of 70 μL after being centrifuged at 300 g for 10 minutes in order to prepare them for analysis. Different markers and antibodies were utilized for cell staining, which included CD31 conjugated with Alexa 647 (BioLegend, 303111), PTPRC antibody conjugated with Alexa 647 (BioLegend®, 304020), Calcein AM (Thermo Fisher Scientific, L3224), and CellTracker™ Orange (CTO) (Thermo Fisher Scientific, C34551), which were added and incubated for one hour. In order to improve the quality of RNA and cell viability, the addition of 15 μL of RPMI containing 10% FBS (Gibco) and 3 μL of RNase inhibitors (Thermo Fisher Scientific, N8080119) was done. After incubation, the staining

reagents were diluted with 13 mL of PBS. After the procedure, the resulting sample was centrifuged at 300 g for 10 minutes, and then the volume was reduced to 45 μL through concentration. To ensure that the CTCs would have ideal floating characteristics in an integrated fluidic circuit (IFC), a cell mixture was prepared by combining 30 μL of a Cell Suspension Reagent (Fluidigm, 101-0434) and 45 μL of CTCs, leading to a total volume of 75 μL for the cell mixture.

3.2.1.4 Integrated fluidic circuit (IFC) operation

To prepare the Polaris integrated fluidic circuit (IFC), the Fluidigm Polaris system was employed to load cell capture beads, prime the circuit, and block the polydimethylsiloxane PDMS channels, which helped prevent the occurrence of non-specific protein absorption/adsorption. Next, during the priming procedure, 48 capture sites were preloaded with beads to form a densely packed bead column, which facilitated the capture and retention of single cells. Following priming, the cell mix containing cells and suspension reagent was loaded into three inlets on the Polaris IFC (25 μL each). Next, capture sites were chosen for single cells exhibiting CTO$^+$ & Calcein AM$^+$ & PTPRC$^-$ & CD31$^-$ characteristics. The single cells were then subjected to mRNA-seq chemistry to generate full-length cDNA and pre-amplification on the IFC through a single-cell processing technique.

3.2.1.5 mRNA-seq library preparation and sequencing

The pre-amplified cDNA was generated using the SMARTer® Ultra® Low RNA Kit, which is designed for Illumina® Sequencing and manufactured by Clontech®. For lysing the selected and sequestered single cells, the Polaris cell lysis mixture was utilized. The mixture for cell lysis is composed of Polaris Lysis Plus Reagent (9.6 μL), Polaris Lysis Reagent (8.0 μL), 3 SMART CDS Primer II A (12 M) (9.0 μL), and Loading Reagent (20X) (1.4 μL). The single-cell lysis process follows a thermal profile of 37°C for 5 minutes, 72°C for 3 minutes, 25°C for 1 minute, and then held at 4°C. The volume of the reverse transcription (RT) mixture is 48 μL. It includes 3.2 μL of Polaris RT Plus Reagent, 10 U/μL SMARTScribe Reverse Transcriptase (100 U/μL), 1 U/μL SMARTer Kit RNase Inhibitor (40 U/μL), 1.2 μM SMARTer Kit SMARTer II A Oligonucleotide (12 μM), 1 mM SMARTer Kit dNTP Mix (10 mM each), 2.5-mM

SMARTer Kit Dithiothreitol (100 mM), and 1X SMARTer Kit 5X First-Strand Buffer, which was used for the reverse transcription of RNA to cDNA. The RT mixture is formulated to match the concentrations of the components in the RT chambers of the Polaris IFC. The RT process involves a thermal protocol that includes 90 minutes of incubation at 42°C (RT), followed by 10 minutes at 70°C for enzyme inactivation, and a final hold at 4°C. The PCR amplification was performed using a 90 μL reaction containing 1X Loading Reagent (20X, Fluidigm, 101-1004), 2X Advantage 2 Polymerase Mix (50X, Clontech, 639206), 0.48-μM IS PCR Primer (12 μM, Clontech, 639206), 0.4-mM dNTP Mix (50X/10 mM, Clontech, 639206), and 1X Advantage® 2 PCR Buffer (10X, Clontech, 639206, Advantage 2 PCR Kit). To initiate the preamplification process, the thermal protocol began with enzyme activation at 95°C for 1 minute, followed by five cycles of 95°C for 20 seconds, 58°C for 4 minutes, and 68°C for 6 minutes. This was followed by nine cycles of 95°C for 20 seconds, 64°C for 30 seconds, and 68°C for 6 minutes, and then seven cycles of 95°C for 30 seconds, 64°C for 30 seconds, and 68°C for 7 minutes. Finally, the process ended with a final extension at 72°C for 10 minutes. The Polaris IFC carrier was used to gather the preamplified cDNAs into 48 distinct outlets. Later, modifications were made to convert the cDNA reaction products into mRNA-seq libraries using the Nextera® XT DNA Sample Preparation Kit from Illumina (FC-131-2004, FC-131-2003, FC-131-2002, FC-131-2001, and FC-131-1096). The pooled library was sequenced on the Illumina NextSeq, and the resulting data were analyzed. To be more specific, the reactions were carried out using a volume equal to 25% of the recommended volume, and the tagmentation step was extended to 10 minutes. Additionally, the PCR step had an increased extension time of 60 seconds instead of 30 seconds. Following the PCR step, the samples were combined and subjected to two rounds of purification using 0.9× Agencourt® AMPure® XP SPRI beads (manufactured by Beckman Coulter) and ultimately eluted using Tris + EDTA buffer. An Agilent high-sensitivity DNA chip was used to determine the amount of DNA. The combined library was sequenced on an Illumina NextSeq® machine with reagent kit v3, producing paired-end reads of 2 × 74 bp.

3.2.1.6 Preprocessing of scRNA-seq datasets

The unCTC R package can process single-cell RNA sequencing data in two different formats: raw count data and transcripts per million (TPM). This chapter employs three datasets, namely, the Ebright et al. dataset and the Ding et al. dataset, which are publicly accessible, and the Poonia et al. dataset, which was generated in-house. The count and TPM data for the Ding et al. dataset were obtained from the single-cell gateway of the Broad Institute, while the count and TPM data for the Ebright et al. and Poonia et al. datasets were acquired by processing their FastQ files. Before processing, the FastQ files underwent quality checks using the FastQC tool (Andrews, 2010). Both datasets were evaluated for their average mean quality score, GC content, and per-sequence quality score using the FastQC tool (Andrews, 2010). The Ensembl (release 75) hg19 reference genome and hg19 GTF file were utilized for aligning the Ebright dataset (Howe *et al.*, 2021). To determine the levels of gene expression, we utilized RNA-seq by Expectation-Maximization v.1.3.1 (RSEM), which involves two scripts, namely rsem-prepare-reference and rsem-calculate-expression (Li and Dewey, 2011). The count profiles and length-normalized TPM datasets, which represented the expression of 57,773 transcripts, were obtained for both studies. Preprocessing steps were performed to prepare the scRNA-seq datasets, and **Figure 3.2** illustrates the details of these steps. To process the Poonia et al. dataset, an index for RNA-seq by Expectation-Maximization (RSEM) was created using hg19/GRCh37, and the RefSeq transcriptome was obtained from the UCSC Genome Browser database (Karolchik *et al.*, 2003). RSEM/Bowtie (Li and Dewey, 2011; Langmead, 2010) was used to align the raw read data to the index, and RSEM v1.2.4 was utilized to quantify the gene expression levels in counts as well as TPM for all genes across all samples. The genomic mappings were calculated by using TopHat2 v2.0.13 (Kim *et al.*, 2013), and the obtained alignments were utilized to determine the percentages of genomic mapping. Using Bowtie 2 v2.2.4 (Langmead and Salzberg, 2012), the raw sequencing read data were aligned directly to the human rRNA sequences NR_003287.1 (28S), NR_003286.1 (18S), and NR_003285.2 (5.8S). Compared to GRCh37, GRCh38 (Guo *et al.*, 2017) altered 8000 nucleotides, negligibly affecting sequences of functional genes. However, the primary claims and conclusions of the study remained unchanged while using GRCh38.

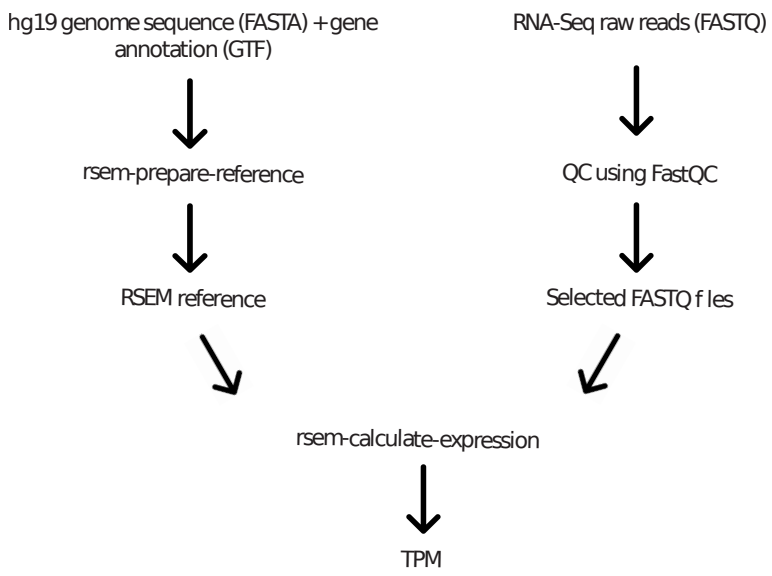

Figure 3.2: Computational workflow for generating transcripts per Million (TPM) Matrix. The computational workflow outlines the basic steps in producing the transcripts per million (TPM) matrix using scRNA-Seq FASTQ files. The genome GRCh38, which has only minimally altered 8000 nucleotides compared to hg19 or GRCh37, was utilized. Despite this variation, the sequences of functional genes remain relatively unaffected.

3.2.2 Processing of CD45+ single cells from a healthy individual

In this study, 7.5 ml of blood (EDTA) was subjected to RBC lysis. The CD45 cells were then labeled with an Alexa647-conjugated CD45 antibody (Biolegend, 304020) and sorted using the Sony SH800 Cell Sorter. The live CD45 cells were subsequently isolated by negatively depleting Zombie Yellow™ Fixable dye-stained (Biolegend, 423103) dead cells using the SH800. The C1 single-cell system (Fluidigm) was used to capture live CD45 cells, and the whole transcriptome library preparation was performed using the same protocol applied to the CTCs. These libraries were sequenced using the Illumina Miseq system with v3 chemistry for 75bp paired-end reads. Out of the 96 cells sequenced, 71 cells with a minimum of 10,000 raw reads were qualified for use as a reference for CNV inference in Poonia et al.'s CTC study.

3.2.3 Integration, filtering, and normalization of data

The RSEM software was utilized to obtain read count and TPM data from the scRNA-seq datasets of Poonia et al. and Ebright et al. We removed cells from the datasets with a total read count of less than 50,000, resulting in the elimination of nine CTCs from

the Poonia et al. dataset **(Table 3.3)**. The Ebright et al. dataset includes 824 cells, and each of the CTCs in this dataset has a read count of at least 50,000. The three datasets were combined based on common genes, followed by filtering steps to remove cells expressing fewer than 1500 genes and genes expressing no expression in fewer than five cells. In the subsequent steps, Single-cell normalization and batch correction were executed using the Linnorm normalization technique with default parameters, encompassing the *Linnorm.Norm* function (Yip *et al.*, 2017). This function performs comprehensive batch effect and library size difference normalization on the input dataset, employing parameters such as minNonZeroPortion,BE_F_p, $BE\_F\_LC\_Genes$, and $BE\_F\_HC\_Genes$, which respectively dictate the minimum non-zero portion threshold for gene inclusion, the filtering criteria based on standard deviation and skewness, and the proportions of lowest and highest expressing genes during the normalization process. The algorithm also incorporates additional parameters like $BE\_strength$, influencing the strength of batch effect normalization, and max_F_LC, determining the maximum threshold assigned for low-expressing genes. The normalized expression values were subjected to log transformation after adding a pseudo count of 1 to the TPM and count matrix.

Table 3.3: Overview of preprocessed CTCs enriched by ClearCell® FX-Polaris™ workflow.

Patient	#CTC	Hormone Receptor Status
P1	11	ER-/PR-/HER2-
P3	10	ER+/PR+/HER2-
P4	12	ER+/PR+/HER2-
P5	32	ER+/PR+/HER2-
P7	4	ER-/PR-/HER2+
P9	3	ER-/PR-/HER2+
Total	72	

3.3 Results

3.3.1 Marker-free capture of CTCs

Out of all types of cancer, breast cancer is the most frequently occurring cancer and is responsible for a significant number of deaths related to cancer (Kamal *et al.*, 2017). Worldwide, breast cancer became the most prevalent cancer in 2020, surpassing lung

cancer. Metastasis accounts for about 90% of breast cancer deaths. CTCs have been extensively studied in breast cancer (Zhang *et al.*, 2021). However, traditional affinity-based methods for detecting CTCs using epithelial markers, such as EPCAM and CK, have limitations because, during the Epithelial-Mesenchymal transition (EMT), these markers are down-regulated. Moreover, using fluorescence-activated cell-sorting techniques is challenging because of the extremely low frequency of circulating tumor cells, which typically amounts to less than one CTC per ten milliliters of blood in non-metastatic cancers (Thery *et al.*, 2019). Because of the limitations of current techniques, there is a need for a reliable method to identify and enrich CTCs in large volumes of blood cells that do not rely on markers. Such methods are attractive because they allow for the detection of a more significant number of CTCs that might be missed due to inconsistent or missing protein marker expression on the surface of CTCs. To address this issue, a marker-free approach was developed by combining the ClearCell® FX and Polaris™ systems (Iyer *et al.*, 2020). As a part of this method, the enrichment of CTCs is done in two steps. First, ClearCell is used for size-based enrichment, followed by the use of Polaris for negative depletion based on CD31 (endothelial cell marker) and PTPRC (leukocyte marker) (Warkiani *et al.*, 2014; Ramalingam *et al.*, 2016) (**Figure 3.3**). Using the ClearCell FX and Polaris systems, 81 individual CTCs were obtained from six female breast cancer patients of three subtypes ($ER^-/PR^-/HER2^+$, $ER^+/PR^+/HER2^-$, and $ER^-/PR^-/HER2^-$), as shown in **Table 3.2**. Out of these, 72 CTCs satisfied the quality control criteria detailed in **Table 3.3**. The following sections detail the unCTC-based analysis of these cells, which differs from three other integration analysis methods: Seurat (Hao *et al.*, 2021), fastMNN (Haghverdi *et al.*, 2018), and Harmony (Korsunsky *et al.*, 2019).

3.3.2 unCTC recognizes CTCs selected by the ClearCell FX and Polaris workflow

Our previous work (Iyer *et al.*, 2020) showed how supervised machine learning methods can be used for CTC characterization. However, since CTC phenotypes are constantly changing, it is important to use unsupervised methods to characterize single CTC transcriptomes. This is especially important when CTCs have atypical phenotypes, as classification-based approaches may not be reliable. To address this limita-

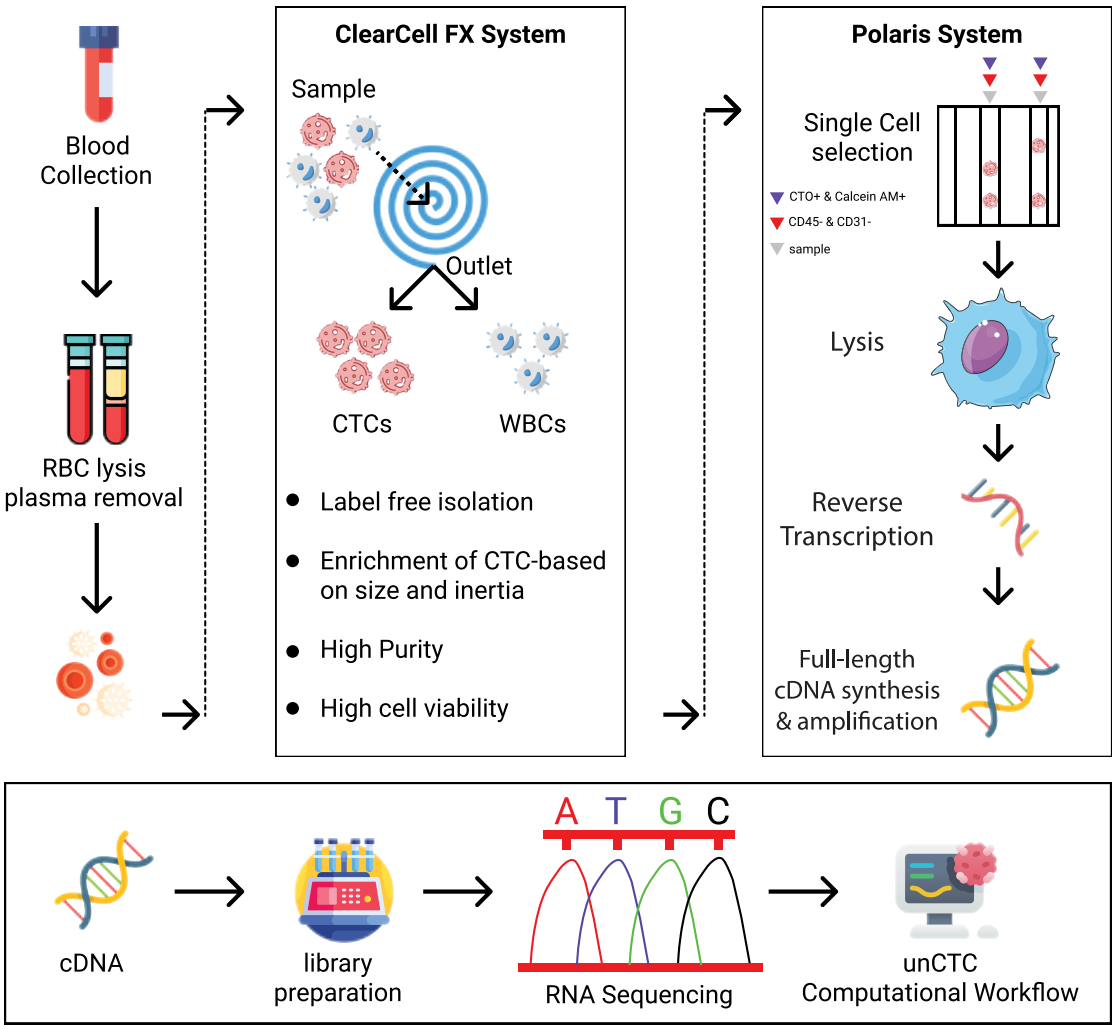

Figure 3.3: The ClearCell® FX and Polaris™ workflow provides a marker-free approach to enriching CTCs. The essential steps of capturing and isolating CTCs involve a two-pronged system, as demonstrated by the schematic diagram. The ClearCell® FX component utilizes a spiral chip to sort CTCs by size, while Polaris™ performs single-cell capture and cDNA synthesis for potential CTCs, followed by the removal of PTPRC/CD31-positive cells. The generated cDNA is then subjected to library preparation and RNA sequencing.

tion, the unCTC approach uses unbiased single-cell characterization tools to analyze CTCs. To validate our findings, we analyzed 72 single-cell transcriptomes captured by the ClearCell® FX and Polaris™ workflow associated with potential circulating tumor cells. These transcriptomes were obtained from six women with three different subtypes of breast cancer: $ER^-/PR^-/HER2^+$, $ER^+/PR^+/HER2^-$, and $ER^-/PR^-/HER2^-$. Additionally, we incorporated a control dataset published in a study by Ebright et al. (Ebright *et al.*, 2020). This dataset contained 824 cells from 45 patients with the ER^+/PR^+ subtype of breast cancer. We analyzed 752 scRNA-seq profiles of WBCs using the Smart-seq2 protocol in two separate runs, as described in Ding et al. (Ding *et al.*, 2020) and presented in **Table 3.1**. **Figure 3.4** shows the capability of four methods, namely Vanilla Seurat, fastMNN, Harmony, and unCTC, to distinguish between WBCs and CTCs. The findings of our study showed that, for the majority of clusters, the fastMNN and Harmony methods identified a combination of CTCs and WBCs **(Figure 3.4D-I)**. Our observations revealed that Vanilla Seurat successfully detected multiple clusters composed solely of CTCs, but there were batch effects present in the embeddings **(Figure 3.4A-C)**. On the other hand, unCTC clustered CTCs into three distinct clusters, while WBCs were grouped into a single large cluster **(Figure 3.4J-L)**. We observed that the ClearCell FX and Polaris workflow selected CTCs clustered with one of the ER^+ subgroups identified in the study by Ebright et al. Out of the 72 CTCs that satisfied the filtering criteria (derived from the ClearCell FX and Polaris workflow), ER^+ cells constituted the largest portion, comprising 54 out of 72. Out of the remaining CTCs, we observed 7 cells that belonged to the $HER2^+$ category and 11 cells that belonged to the triple-negative category. One explanation for not being able to distinguish $HER2^+$ and triple-negative CTCs as distinct categories could be due to their insufficient numbers; this could pose a difficulty for the unCTC method to maintain important genes and pathways throughout various upstream filtering stages, such as pathway selection and gene filtering.

Furthermore, we examined Integrative Seurat utilizing RPCA and CCA variations. However, Integrative Seurat was unable to distinguish between CTCs and WBCs, as it combined all of them into all 13 cluster clusters. **Figure 3.5** displays the UMAP scatter plot for CCA **(Figure 3.5A-C)** and RPCA **(Figure 3.5D-F)**, which allows for an insightful visualization of the enriched CTCs through ClearCell® FX-Polaris™ workflow. To ensure an unbiased assessment of the clustering quality, we applied three

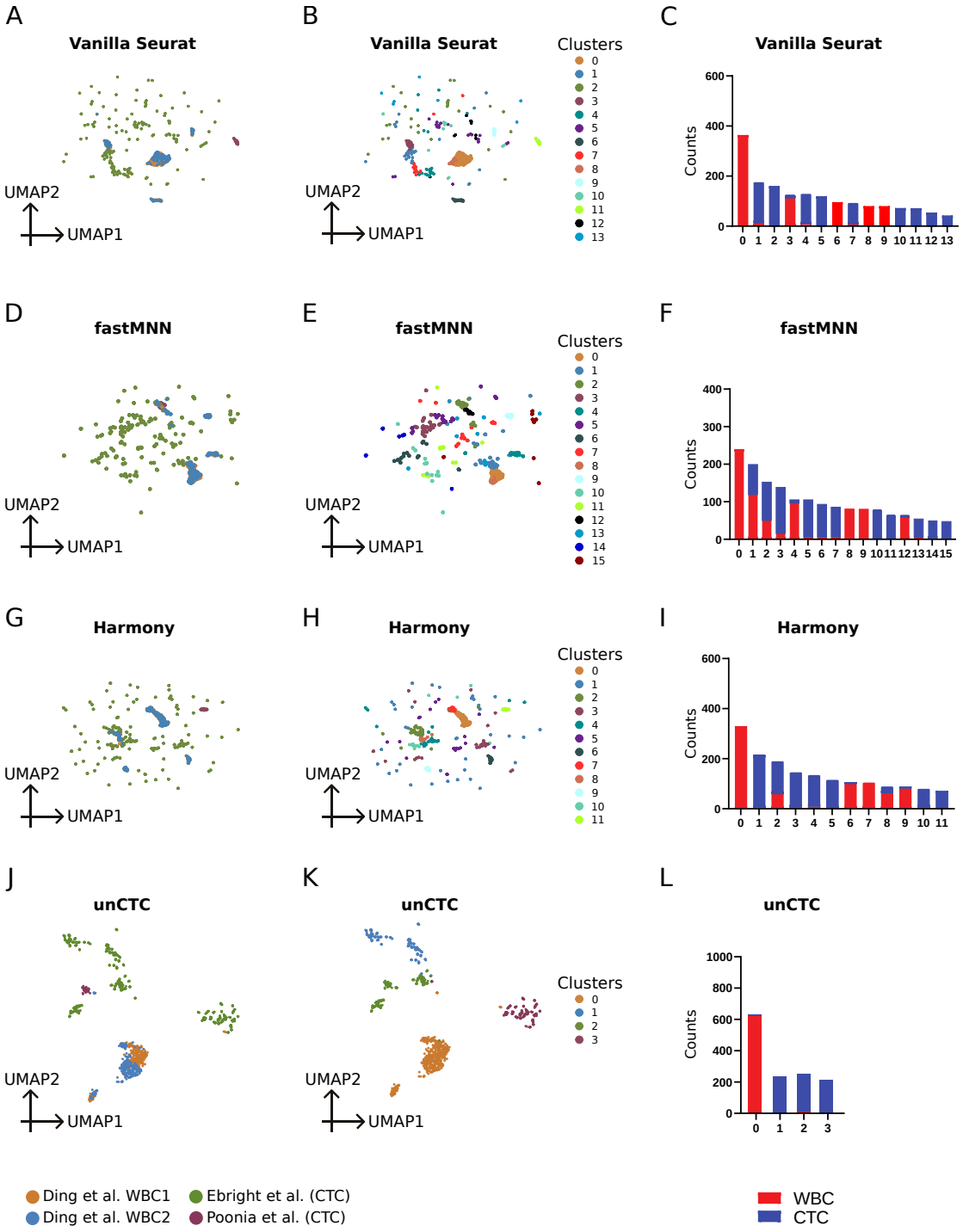

Figure 3.4: CTC Clustering Exploration in ClearCell® FX-Polaris™ System. The clustering of CTCs obtained from the ClearCell® FX-Polaris™ system was evaluated using various methods (A-I). Cluster purity was assessed using independent breast CTC and WBC scRNA-seq profiles, compared across Vanilla Seurat, fastMNN, and Harmony (J-L). Notably, Vanilla Seurat, fastMNN, and Harmony failed to integrate CTCs from different sources, while unCTC accurately segregated CTCs and WBCs. Additionally, CTCs from the ClearCell® FX-Polaris™ system co-clustered with breast CTCs from Ebright et al. (Ebright *et al.*, 2020).

widely accepted metrics - cluster purity, Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). Impressively, our analysis demonstrated that the unCTC method outperformed the current techniques in all three metrics, signifying its superior performance **(Figure 3.5G)**. Besides analyzing the TPM data used in this study, we also evaluated the scRNA-seq count profiles of the corresponding study with three integrative analysis methods, namely Vanilla Seurat, Harmony, and fastMNN. Nonetheless, our newly developed method, unCTC, outperformed all of these methods by successfully distinguishing CTCs and WBCs into separate clusters. To visually represent the data, we generated 2D scatter plots of cells using UMAP, which can be viewed in **Figure 3.6**.

3.3.3 Marker-dependent characterization of CTC clusters

Understanding cell lineages requires the enrichment of lineage markers that are well characterized. To achieve this objective, two methods can be employed: analyzing the distinct expression of individual markers and analyzing marker panels. Our study focused on exploring the lineage identities of the clusters that were identified by DDLK, as shown in **Figure 3.7A**. We identified several widely recognized immune-cell markers, including *NKG7, PTPRC, PTPRCAP, IL32, CD74*, and *CD48* among the top 200 genes that were differentially up-regulated in cluster 0, which mainly comprised white blood cells from the Ding et al. dataset. Supplementary Table S5 in Poonia et al. study provides further details (Poonia *et al.*, 2022). Clusters 1, 2, and 3 mainly comprised CTCs from the Poonia and Ebright et al. datasets. Integrins (*ITGA2B* and *ITGB5*) expression levels were observed to be elevated in Cluster 1 among the identified clusters. Adhesion molecules called Integrins play a pivotal role in hemostasis and platelet function. Based on RNA extracts from single and clustered CTCs, some recent studies have proposed a potential interaction between CTCs and platelets (Aceto, 2020; Szczerba *et al.*, 2019; Ting *et al.*, 2014). The molecular profiles of CTCs are affected by their continuous interaction with blood components like platelets, extracellular vesicles, and circulating nucleic acids (Ward *et al.*, 2021). We noted upregulated levels of *CLU* and *SPARC* in our analysis, which are markers for platelet degranulation and have been reported to regulate *PF4*. *PF4* is an important endocrine factor that has been previously linked to adverse outcomes in patients with lung cancer. Cluster 1 specific CTCs exhibited

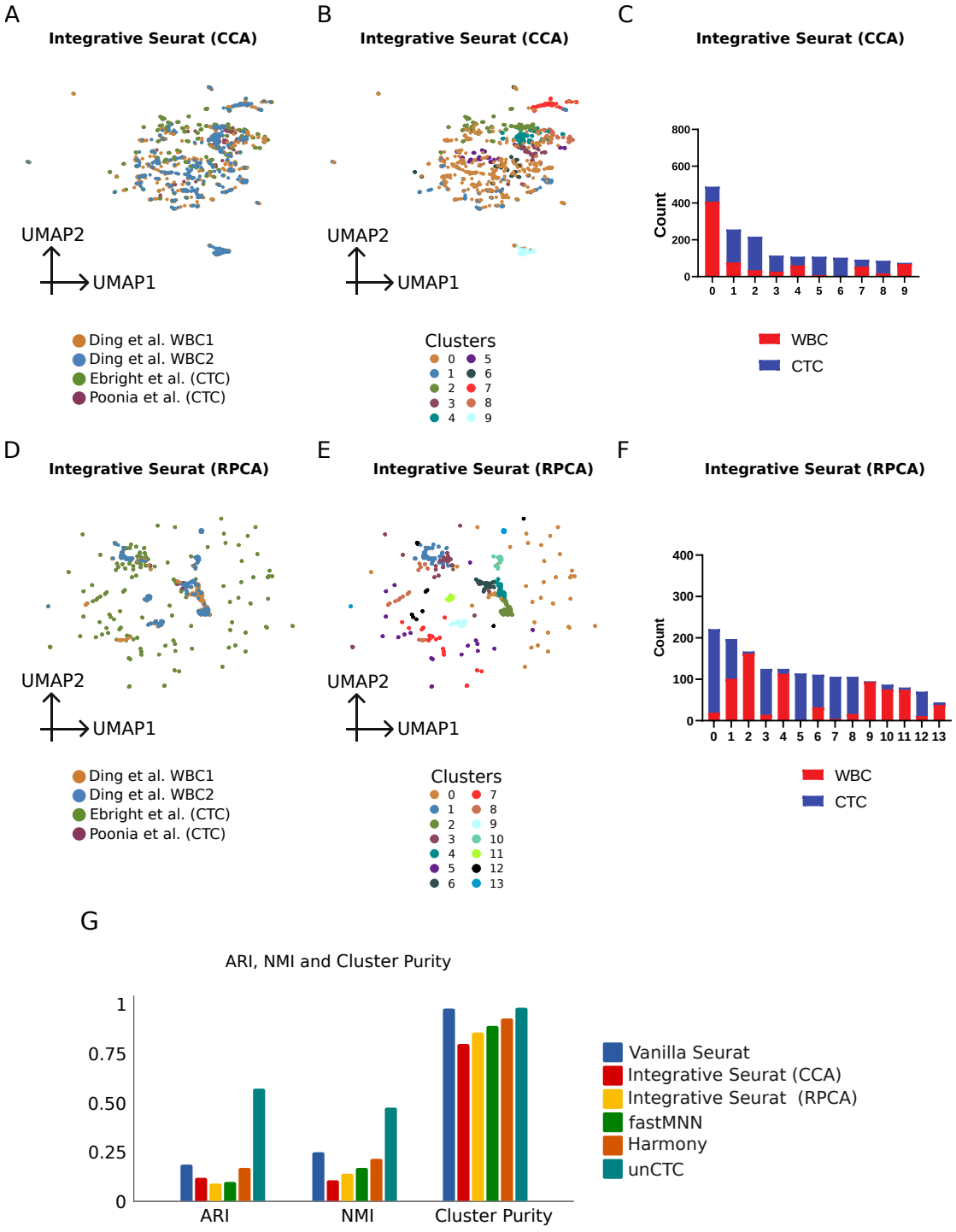

Figure 3.5: Seurat's integrative analysis. (A-C) Illustrate the use of Seurat's canonical correlation analysis (CCA) and (D-F) illustrate the reciprocal principal component analysis (RPCA) methods for integrating multiple datasets from different studies, including the Poonia dataset, Ebright dataset, and Ding dataset. (G) The performance of various methods used in this study is summarized based on a comparison of clusters with WBC/CTC annotations.

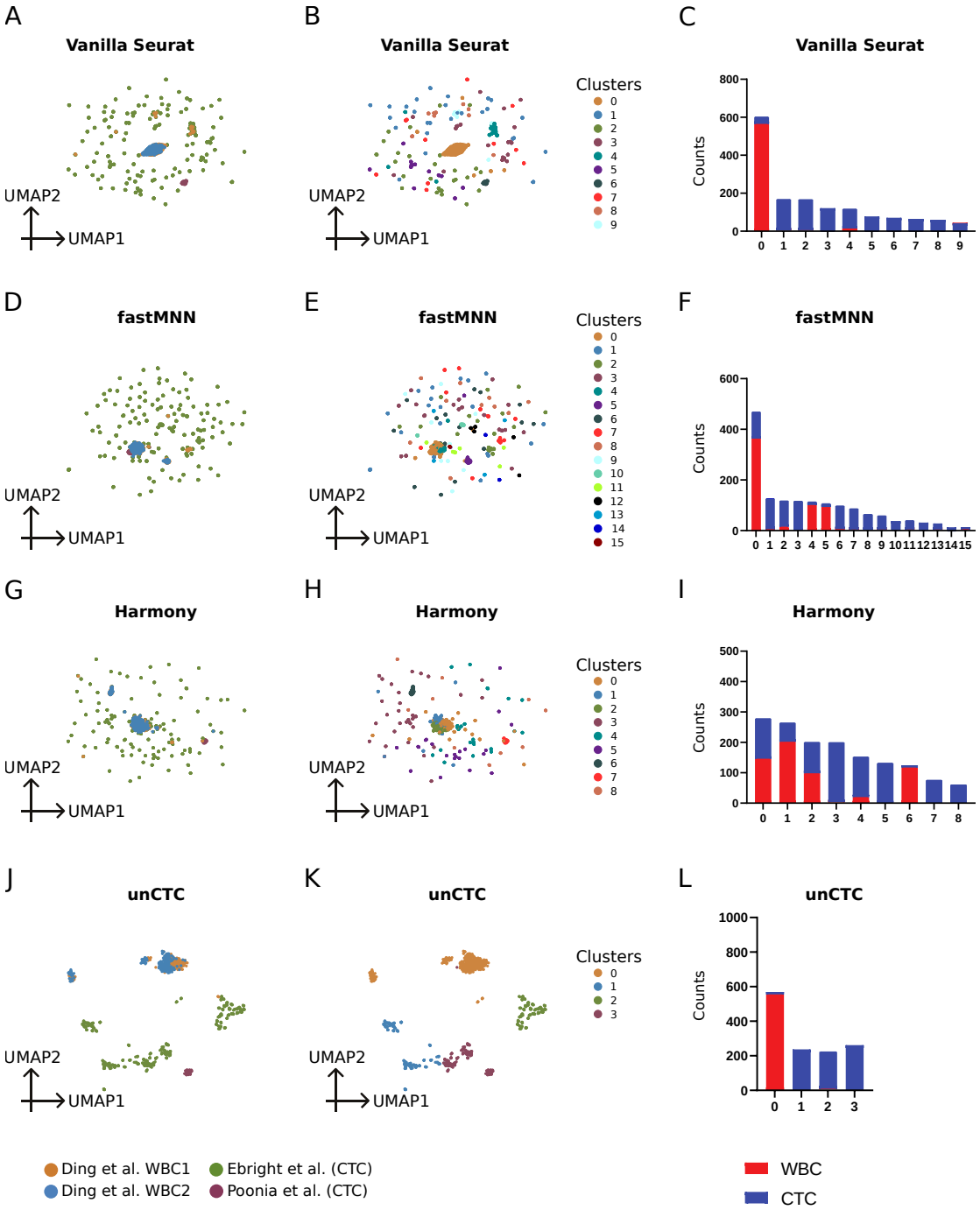

Figure 3.6: The scRNA-seq expression count profile of CTCs obtained from the ClearCell® FX - Polaris™ system was analyzed using clustering and UMAP-based visualization. The data was processed with various methods including Vanilla Seurat, fastMNN, Harmony, and unCTC. The resulting figures (A-I) show the clustering and distribution of scRNA-seq profiles for CTCs and WBCs using Vanilla Seurat, fastMNN, and Harmony. Figures (J-L) show equivalent results using unCTC.

elevated expression of several oncogenes that play important roles in the progression of breast cancer. These included *CDKN1A, TIMP1*, and *PGRMC1*, which have been extensively studied in the literature (Abreu *et al.*, 2020; Clark *et al.*, 2016; Koch *et al.*, 2020). Cluster 2, which contained CTCs from ClearCell/Polaris as well as CTCs from the Ebright dataset, displayed elevated expression levels of transcripts associated with breast cancer. These transcripts consisted of *POU5F1*, a marker for cancer stem cells (Jin *et al.*, 2019*b*); *BRIP1*, whose genetic mutation increases the risk of breast cancer and its high expression is associated with the invasive primary disease (Eelen *et al.*, 2008); *IL10*, which has been reported to enhance breast cancer proliferation and progression (Sheikhpour *et al.*, 2018); and *IDO1*, an essential protein for immune checkpoint regulation (Dill *et al.*, 2018). Cluster 3 was identified by higher expression levels of typical epithelial markers such as *KRT18, EPCAM*, and *KRT19*. Moreover, there was an upregulation of the tumor suppressor *SOD1* in this cluster (Liu *et al.*, 2020).

3.3.4 Concordance of cluster-specific pathway enrichment with differential gene expression analysis in clusters

The up-regulated genes specific to each cluster were subjected to functional analysis using IPA (QIAGEN Inc.) (Krämer *et al.*, 2014). This strategic decision to focus exclusively on upregulated genes and pathways serves the purpose of streamlining our analysis, directing attention to biologically relevant molecular signatures that actively contribute to the distinctive characteristics of each cluster. Our study revealed that the up-regulated genes in Cluster 0, which represents the WBC cluster, were largely not associated with cancer. On the other hand, the other clusters, which were specific to CTCs, demonstrated a substantial enrichment of pathways associated with cancer. This result was consistent with our previous analysis of marker genes for each cluster **(Figure 3.7B-E)**. Furthermore, we created visual representations of certain relevant pathways that exhibited a notable increase in enrichment in particular clusters **(Figure 3.8)**. Our observations regarding the differential enrichment of pathways specific to each cluster revealed noteworthy trends. Notably, in Cluster 0, which comprises cells from the Healthy White Blood Cell (WBC) class, we observed an upregulation of blood-related pathways, such as "T helper cell pathways," "Cytotoxic T cell pathways," "Monocyte pathways," and various "Lymphocyte" pathways. Conversely, Clusters 1, 2,

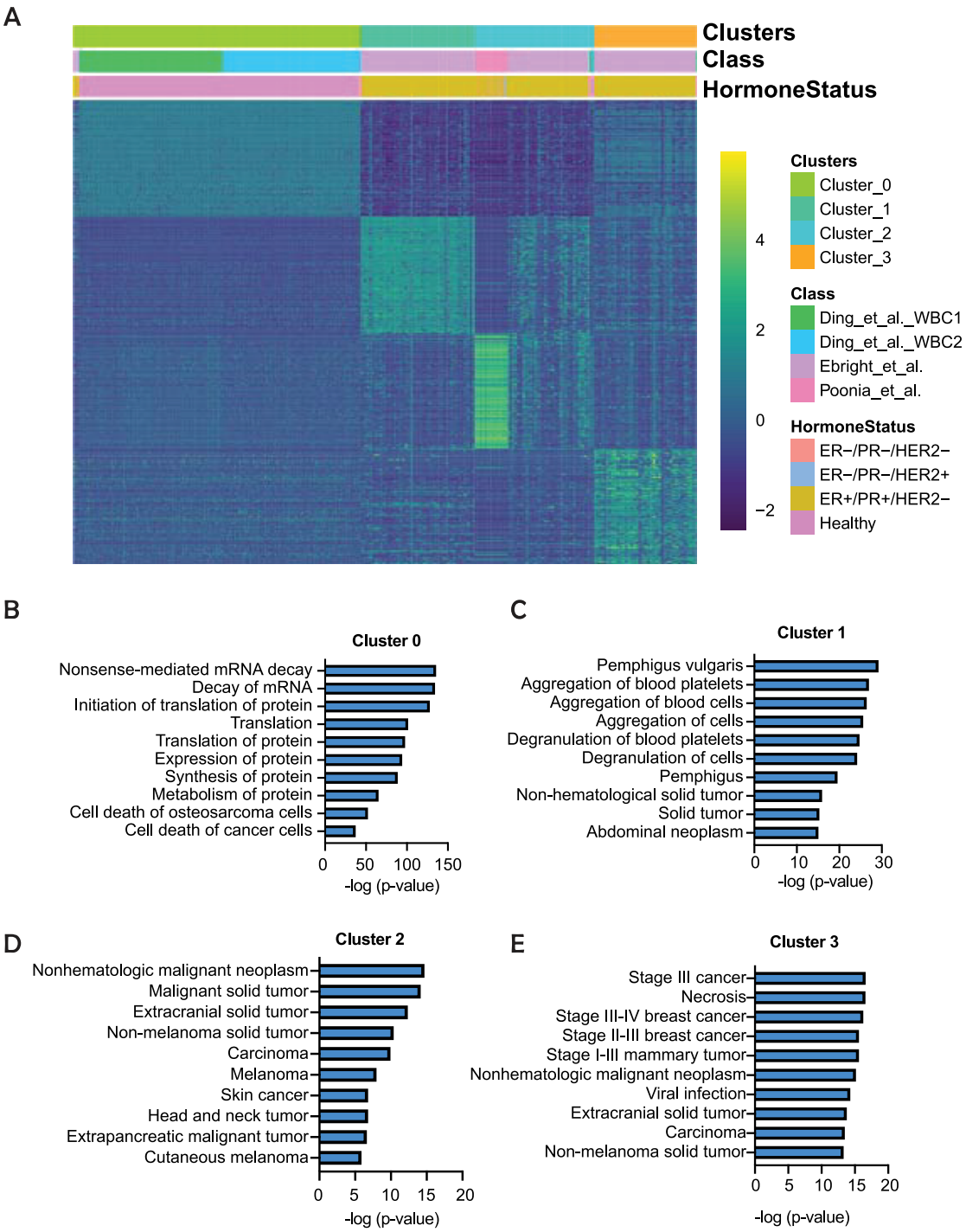

Figure 3.7: Functional enrichment of genes associated with CTCs and WBCs. In panel (A), the heatmap illustrates the expression patterns of the top 200 upregulated genes in each of the four identified clusters, determined using the Limma voom package in R (Law *et al.*, 2014). The color bars in the heatmap signify cluster identity, data source, and molecular subcategories. To unravel the functional implications of these upregulated genes, gene set enrichment analysis was performed using the IPA software. Specifically, the top 200 upregulated genes from each cluster were employed for this analysis. Bar plots (B-E) present the outcomes of the enrichment analysis, with each plot corresponding to a distinct cluster, providing a nuanced understanding of the enriched functional annotations. Additional information on the differentially expressed genes for each cluster can be found in Supplemental Table S5 of the Poonia et al. study (Poonia *et al.*, 2022).

and 3 exhibited upregulation of cancer-related pathways, with a specific emphasis on breast cancer pathways, as the cells in these clusters belong to the breast cancer class. These findings were consistent with the results obtained from the analysis of genes exhibiting differential expression. By incorporating these pathway analyses, we were able to gain a deeper understanding of the functional implications of our results, which can aid in the development of targeted interventions and personalized therapies.

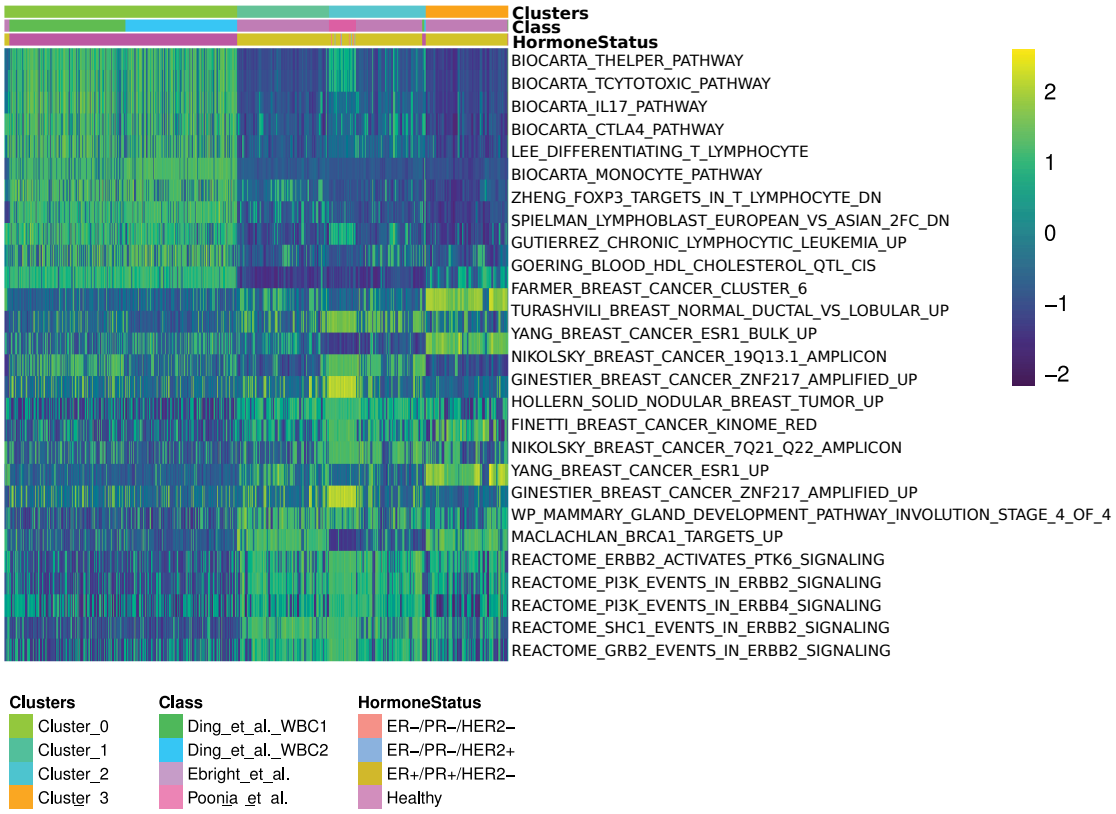

Figure 3.8: Differential pathway analysis. A heatmap was created to display the differential pathway enrichment scores of four clusters identified by unCTC. The color bars represent the identity of the clusters, source data information, and molecular subcategories.

3.3.5 Spatial segregation of CTCs of the TNBC subcategory

The ClearCell/Polaris CTCs were grouped with a subset of Ebright CTCs based on the clustering analysis. It's important to note that all Ebright CTCs belong to the $ER^+/PR^+/HER2^-$ category. Additionally, a significant proportion of Poonia CTCs (54 out of 72) also belong to this category. As previously discussed, cluster 2 CTCs exhibited an enrichment of breast cancer markers, which may be the primary reason

behind the clustering of Poonia CTCs with a portion of Ebright CTCs. This clustering has the potential to mask the inter-subtype heterogeneity that is present in Poonia CTCs. To better understand this heterogeneity, we conducted a specific analysis using unCTC on Poonia CTCs. This examination showed that there were two distinct groups of $ER^+/PR^+/HER2^-$ CTCs and also revealed the spatial distribution of triple-negative breast CTCs **(Figure 3.9A)**. The dataset of Poonia et al. that employs the ClearCell FX/Polaris worflow encompasses all the critical subtypes of breast cancer, including $ER^+/PR^+/HER2^-$, $ER^-/PR^-/HER2^+$, and $ER^-/PR^-/HER2^-$. We detected the top 10 genes that were up-regulated for each subtype **(Figure 3.9B)**. We observed an increased expression of the *HEYL* gene in the $ER^-/PR^-/HER2^+$ subtype. It's important to mention that the *HEYL* gene's overexpression is linked to unfavorable survival outcomes in estrogen-negative breast cancer, as reported earlier (Han *et al.*, 2008). A different gene that is significant, *CREBL2*, has been discovered to exhibit higher expression levels in a cell line with $ER^+/PR^+/HER2^-$ subtype compared to a cell line with the basal subtype (Mellick *et al.*, 2002). In our analysis, we also found an up-regulation of *CREBL2* in $ER^+/PR^+/HER2^-$ cells. Furthermore, we noticed increased levels of expression of *EIF3C* in CTCs that originated from the Triple Negative Breast Cancer (TNBC) subtype. As previously reported, this gene is recognized for its ability to stimulate cell proliferation (Zhao *et al.*, 2017).

3.3.6 Cluster characterization using individual and combinatorial markers

We conducted cluster characterization using individual markers and panels of markers. It is widely recognized that the loss of epithelial properties in CTCs leads to a low proportion of CTCs that exhibit typical epithelial markers (Iyer *et al.*, 2020). Since CTCs undergo epithelial-mesenchymal transition (EMT), and scRNA-seq data has a high dropout rate, depending solely on individual markers may not yield substantial differential expression. Therefore, using a combination approach may be more beneficial in monitoring the diverse expression patterns of genes. We selected markers from existing literature that are expressed to a significant degree in immune cells and breast epithelia. We used Stouffer's method to assess the enrichment of combinatorial markers at a single-cell level (Stouffer *et al.*, 1949). This approach allowed us to separate

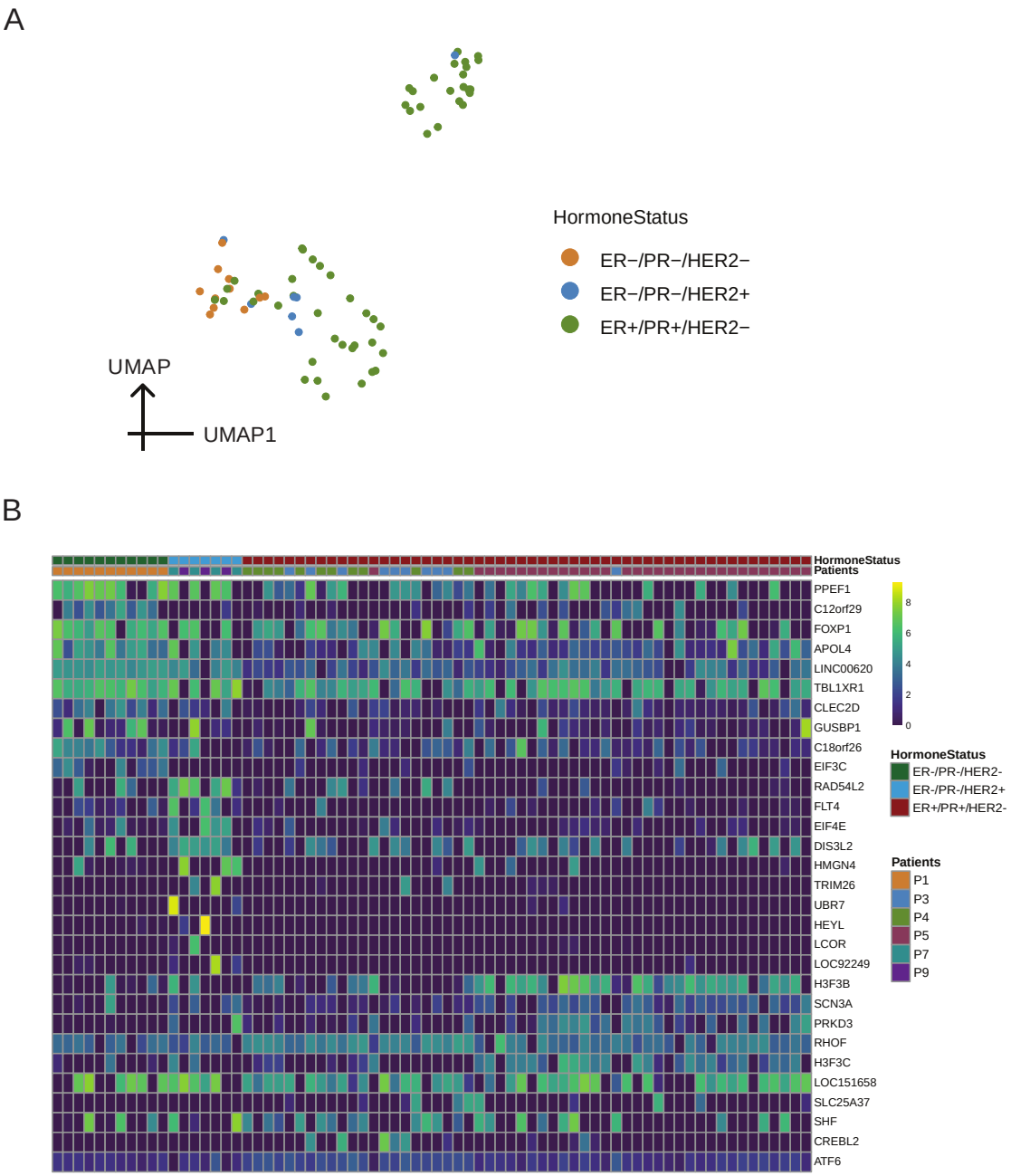

Figure 3.9: The RNA-seq count data of Poonia et al. was analyzed to identify different subpopulations of CTCs. (A) UMAP projection was created based on pathway scores, which showed two subpopulations of CTCs with $ER^{+}/PR^{+}/HER2^{-}$ and spatially segregated TNBCs. Differential gene expression was also analyzed using the Limma package (Ritchie *et al.*, 2015) and its voom (Law *et al.*, 2014) method with default parameter settings. (B) The resulting heatmap displays the gene expression for ClearCell® FX and Polaris™ selected CTCs across three molecular subtypes.

the WBC and CTC populations as expected, as illustrated in **Figure 3.10A**. We then proceeded to track the differential expression of single markers. Specifically, we found that Cluster 0 specific cells expressed elevated expression of leukocyte markers, including *PTPRC* and *NKG7*. Conversely, cells specific to Clusters 1, 2, and 3 displayed comparatively higher expression levels of *EPCAM* and *KRT18* (**Figure 3.10B-E**).

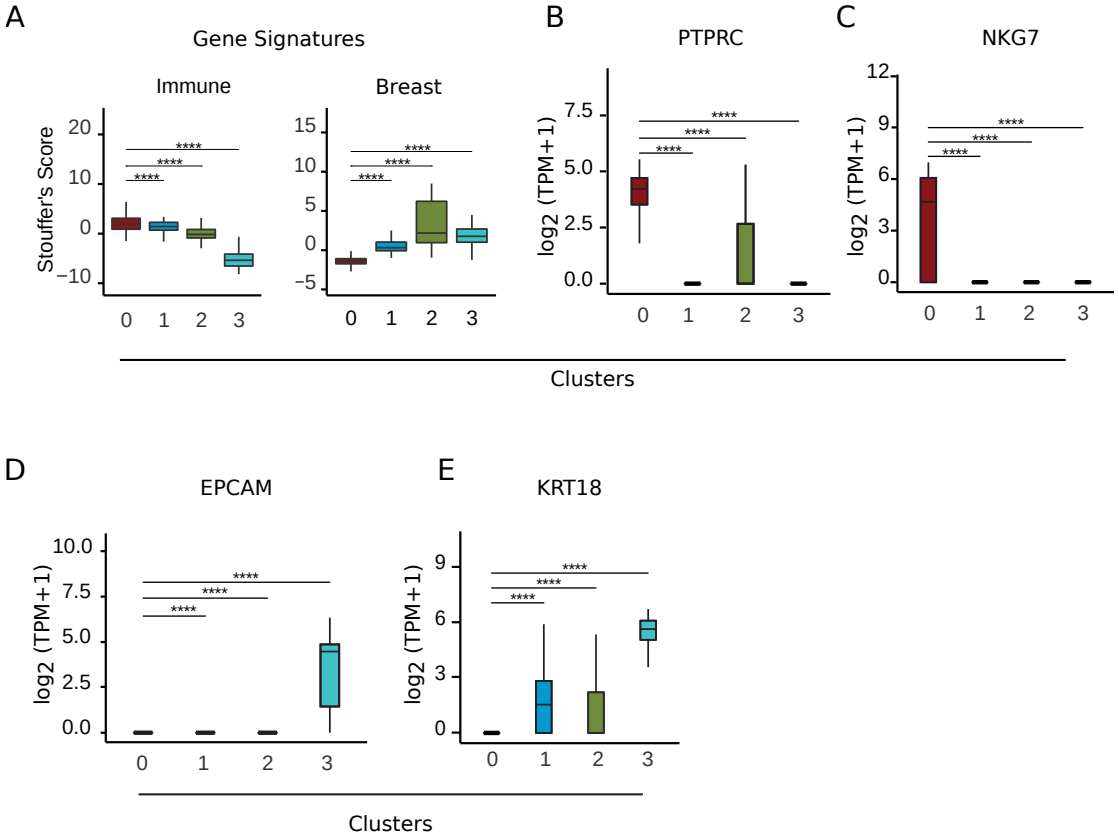

Figure 3.10: Exploring gene set enrichment: Insights from Stouffer's scores and specific markers in single-cell clusters.(A) The distribution of Stouffer's scores (Stouffer *et al.*, 1949) associated with genesets specific to immune cells and breast epithelia was analyzed using a box plot. The enrichment of immune cell-specific markers was observed in Cluster 0, while the other clusters showed enrichment of markers specific to breast epithelia. Two box plots (B and C) show the difference in enrichment levels for two immune cell markers, *PTPRC* and *NKG7*, respectively, across different clusters. The other two box plots (D and E) depict the differential enrichment of two epithelial markers, *EPCAM* and *KRT18*, respectively, across the same clusters.

3.3.7 Expression-based copy number variation inference of CTCs obtained by ClearCell FX and Polaris workflow

Copy number variations (CNV) are genomic alterations characterized by the duplication or deletion of significant chromosomal regions. These somatic copy number variations (CNVs), which are also known as copy number aberrations (CNAs), are commonly detected in cancer and have been associated with its onset, development, and metastasis (Jiang *et al.*, 2016; Urrutia *et al.*, 2018; Sudmant *et al.*, 2015). Studies such as the International Cancer Genome Consortium (Mafficini and Scarpa, 2018) and the Cancer Genome Atlas (Weinstein *et al.*, 2013) have demonstrated their prevalence in various types of cancer. Through single-cell sequencing, expression-based CNV inference is becoming increasingly important in the marker-independent identification of malignant cells within the tumor microenvironment (Tickle *et al.*, 2019). This approach also applies to the circulating tumor cells characterization. To analyze the circulating tumor cells from Poonia et al. and Ebright et al., we used inferCNV, which is a tool for inferring copy number variations (CNVs) from single-cell expression data (Tickle *et al.*, 2019). We performed distinct analyses for each dataset since they were obtained using different experimental methods and chemistries. Additionally, inferCNV does not incorporate a correction for batch effects. To serve as a reference for Poonia et al. CTCs, we utilized internal CD45 cells from a healthy individual, while for Ebright et al. CTCs, we utilized a dataset referred to as the Xu et al. dataset as a control. Xu et al. dataset contains 1000 PBMCs that were selected randomly from a larger group of 7121 PBMCs that were separated using density-gradient centrifugation. All data were used in raw count form for InferCNV. The InferCNV plots revealed significant CNVs across CTCs compared to the reference scRNA-Seq profiles for both datasets. To pinpoint the location of a particular chromosome gain or loss in CTCs, we relied on cytoband information derived from GRCh37 (Barrios and Prieto, 2017). This allowed us to identify the specific p or q-arm positions on the chromosome where the alteration had occurred. We observed similar CNV patterns in both Poonia et al. and Ebright et al. CTCs at specific chromosomal sites including Chr19q13.43, Chr19q13.2, Chr17p13.3, Chr16q24, Chr5q35, Chr4p16.3, Chr3q29, Chr1q21.3, Chr1p36.11 and Chr1p36 (**Figure 3.11 and Figure 3.12**). Breast cancer development has been associated with the 1q chromosome, which harbors both oncogenes and tumor suppressor

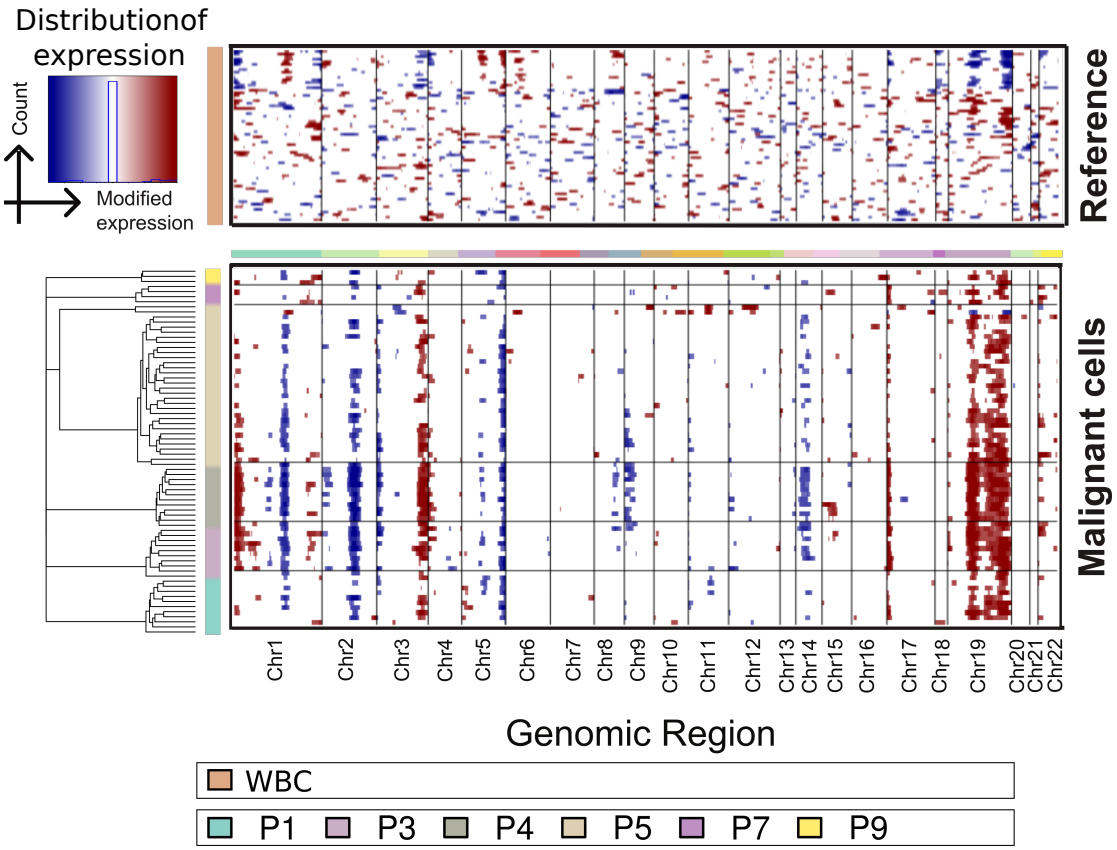

Figure 3.11: Expression-based inference of the CNV landscape across patient-wise malignant cells(Tickle *et al.*, 2019). The heatmap, generated using the inferCNV tool (Tickle *et al.*, 2019), illustrates the putative CNV landscape of circulating tumor cells obtained from six breast cancer patients (p1, p2, p3, p5, and p9), with healthy WBCs serving as the reference. The colored histograms in the upper left visually depict the distribution of gene expression, unveiling the spread of expression values across various samples.

genes (Bièche *et al.*, 1995). Earlier research has identified two altered areas on chromosome 1q: the most frequently deleted small region and the regions that are overrepresented at 1q41-q44 and 1q21-31 (Lobo, 2008; Privitera *et al.*, 2021; Bièche *et al.*, 1995). A potential biological marker for detecting local relapse in breast cancer could be the loss of specific alleles in the chromosomal region 1q21-23 (Gaki *et al.*, 2000; Salahshourifar *et al.*, 2015). Frequent observation of loss of heterozygosity at 1q21.3 has been found to be crucial in the malignancy of tumors, which can serve as a genetic marker for malignancy diagnosis (Salahshourifar *et al.*, 2015; Yang *et al.*, 2005). Alterations on the chromosome are commonly observed in the 1p36 region, which is a well-established hotspot (Ragnarsson *et al.*, 1999). The *TP73* gene is found on the chromosomal region 1p36 and is believed to have tumor suppressor properties, similar to *TP53* (Garnis *et al.*, 2005). An association has been found between amplification in

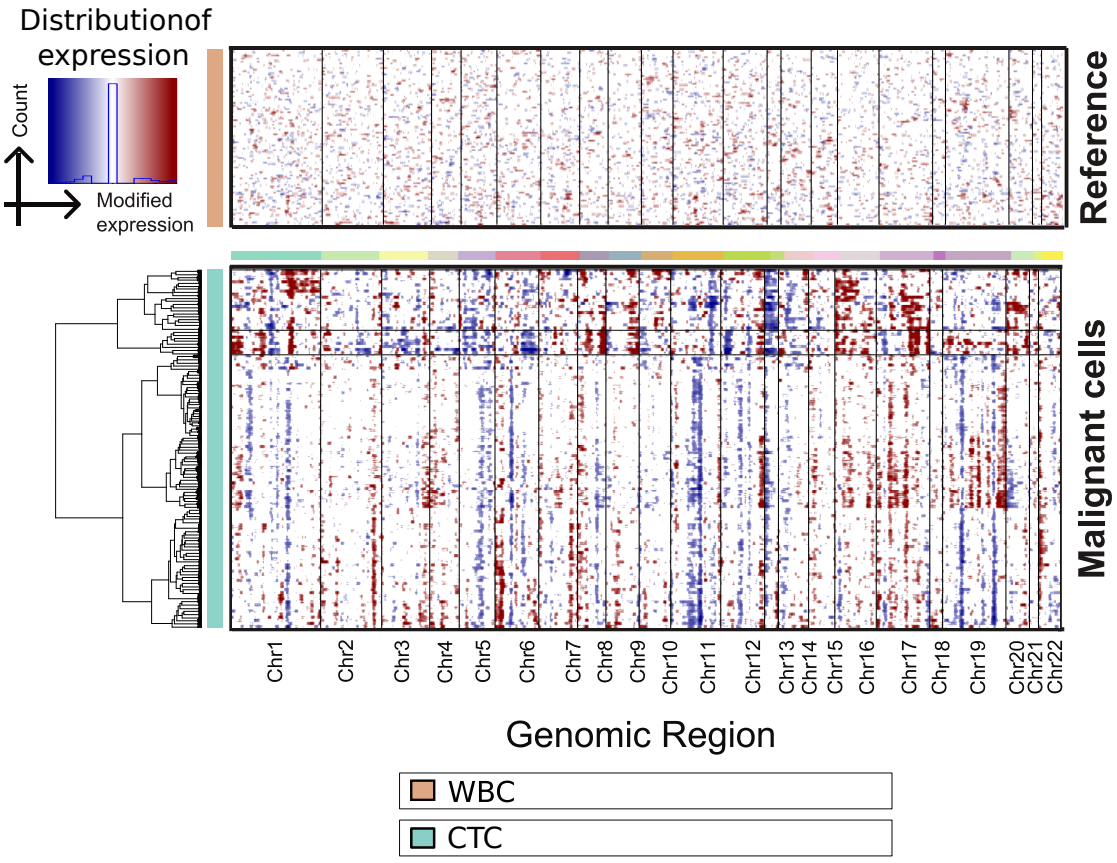

Figure 3.12: Exploring Copy Number Variation (CNV) in Circulating Tumor Cells. The inferCNV tool (Tickle *et al.*, 2019) was used to generate a heatmap of copy number variation (CNV) in CTCs from the Ebright et al. dataset (Ebright *et al.*, 2020). The Xu et al. dataset (Xu *et al.*, 2015) of peripheral blood mononuclear cells (PBMCs) was used as a reference dataset for CNV analysis. The colored histograms in the upper left visually depict the distribution of gene expression, unveiling the spread of expression values across various samples.

the chromosomal region 1p36.33 and unfavorable clinical outcomes (Ragnarsson *et al.*, 1999; Bhosale *et al.*, 2017). Previous studies have documented that there are limited genomic gains that resemble those found in CTCs on Chromosome 19 in primary breast cancer (Kanwar *et al.*, 2015). Additionally, Chromosome 19 gains have been associated with aggressive breast cancer in several studies (Natrajan *et al.*, 2012; Turner *et al.*, 2010). The 19q13 region is linked to the copy number gains of signatures that promote tumor aggressiveness and dormancy in CTCs. *AKT2* (19q13.2), *PAK4* (19q13.2), and *CEBPA* (19q13.11) are genes involved in the promotion of metastasis, invasion, and epithelial-to-mesenchymal transition (EMT) (Kanwar *et al.*, 2015). The *APOBEC3B* gene is positioned on Chromosome 22q13.1 and is involved in regulating ER transcriptional activity, which influences breast cancer cell development (Cescon *et al.*, 2015). In ER^+ breast cancer patients, poor survival rates have been associated with copy num-

ber gain of *APOBEC3B* (Murakami *et al.*, 2021; Periyasamy *et al.*, 2015). Poonia et al. (2022) provide a summary of every detected event, along with source data identities, in Supplemental Table S8 and S9 (Poonia *et al.*, 2022).

3.4 Discussion

In this study, we applied the unsupervised clustering algorithm used in unCTC to Circulating Tumor Cells obtained based on size and marker-based white blood cell (WBC) depletion. The objective was to detect and characterize CTCs using the unCTC pipeline. We employed both count data and transcripts per million (TPM) normalized data and reported the results in Chapters 2 and 3. To compare the integrative analysis performance of the methods, we compared the unCTC method with Seurat (Hao *et al.*, 2021), fastMNN (Haghverdi *et al.*, 2018), and Harmony (Korsunsky *et al.*, 2019). Results showed that unCTC consistently outperformed the other methods. Vanilla Seurat also produces adequate segregation of the target cell types but exhibits batch effects in the embeddings **(Figure 3.4A-C)**. The unCTC method clustered CTCs into three distinct clusters, while WBCs were grouped into a single large cluster **(Figure 3.4J-L, Figure 3.6)**. Despite the fact that unCTC has shown resilience towards batch effects, there is still room for enhancement in accurately interpreting the heterogeneity of cancerous cells. The current study is one of the few to tackle the unique challenge of CTC transcriptome analysis, and it presents a unified computational framework and datasets that can serve as a benchmark in this field of study.

In the field of bioinformatics, it is important for integrative analysis methods to not be influenced by pre-normalization strategies. This is a significant challenge that needs to be addressed. Our study findings indicated that, out of all the methods evaluated, only unCTC demonstrated consistent and equitable performance in all cases, including count and TPM datasets. This performance was evident in the ability of unCTC to distinguish between white blood cell (WBC) and circulating tumor cell (CTC) subpopulations, as well as its capacity to generate cohesive cell embeddings.

Our study has identified precise genomic locations that indicate amplifications or deletions which have been previously well-established in breast cancers. The validity of our findings was further supported through the use of expression-based copy number

variation (CNV) inference. This analysis revealed analogous CNV patterns in CTCs from both Ebright et al. and Poonia et al. studies, particularly at specific chromosomal sites. The implications of these findings are noteworthy since they propose that such chromosomal sites could be distinctive attributes of breast cancer and might be utilized as targets for therapy in the treatment of breast cancer.

Our analysis of the CTC dataset from Poonia et al. using unCTC revealed spatial segregation of CTCs belonging to the TNBC subcategory. Detecting TNBCs is difficult because they do not have canonical surface markers. As a result, many studies on circulating tumor cells concentrate on other types of breast cancer. Our study suggests that the ClearCell® FX - Polaris™ system has the potential to accurately detect CTCs in triple-negative breast cancer patients due to its marker-independent methodology. It is important to mention that Ebright and colleagues utilized the CTC-iChip technology (Ozkumur *et al.*, 2013) for capturing CTCs and then utilized epithelial or other cancer-specific markers for identifying CTCs after the enrichment process. However, this approach may not be suitable for detecting TNBCs since surface markers are not present in this type of breast cancer.

In summary, the utilization of unCTC as an integrative analysis method is a highly promising development that has the potential to revolutionize our understanding of CTCs and their significance in cancer diagnosis, prognosis, and treatment. It is crucial to have an integrative analysis method that is independent of pre-normalization strategy, and unCTC was the sole technique that demonstrated consistent and unbiased performance across all scenarios, thereby making it a effective method for distinguishing between WBC/CTC subpopulations and generating cohesive cell embeddings. Therefore, the use of unCTC can pave the way for further progress in the analysis of CTCs and ultimately lead to better patient outcomes.

3.5 Data access

The sequencing data obtained from this research, both in raw and processed form, have been uploaded to the NCBI Gene Expression Omnibus (GEO) database. Accession numbers for the data are available at the following links: GSE186288 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE186288), and

GSE210651 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE210651).

CHAPTER 4

Identifying pre-transplant risk factors associated with early relapse in multiple myeloma patients undergoing autologous stem cell transplantation

4.1 Introduction

Multiple Myeloma (MM) is a type of cancer that affects plasma cells, characterized by the expansion of malignant plasma cells in the bone marrow and the detection of monoclonal protein (M-protein) in the blood and urine (Kyle and Rajkumar, 2004; Palumbo and Anderson, 2011). It is the second most prevalent form of hematologic cancer after non-Hodgkin lymphoma (Kazandjian, 2016). MM accounts for approximately 15-20% of deaths related to hematologic malignancies and roughly 2% of all cancer-related deaths (Smith and Yong, 2013).

The latest estimates by the American Cancer Society project around 35,730 new cases of multiple myeloma in the United States in 2023, with 19,860 cases in men and 15,870 cases in women. Additionally, it is estimated that there will be approximately 12,590 deaths due to multiple myeloma, with 7,000 deaths in men and 5,590 in women (Siegel *et al.*, 2023; Fitzmaurice *et al.*, 2017; Parkin *et al.*, 2005). Over the past decade, multiple myeloma has risen in rank from 23rd to 21st place among the 32 different types of cancer studied worldwide (Fitzmaurice *et al.*, 2017). The lifetime risk of developing multiple myeloma in the US is 1 in 132 or 0.76% (Varshney *et al.*, 2022; Bethesda, 2018). Furthermore, studies have shown that black people have twice the incidence of multiple myeloma compared to white people (Waxman *et al.*, 2010).

The five-year survival rate for multiple myeloma patients has consistently improved over the past twenty years. It increased from 25% (1975-1977) to 47% (2004-2010) (Siegel *et al.*, 2015). Currently, the medical community observes a five-year survival rate of 50% for individuals diagnosed with multiple myeloma (Wong and Tay, 2018;

Cancer.org, 2018). Although there is no known cure for multiple myeloma, advancements in medical treatments such as chemotherapy, stem cell transplantation, and immunotherapy have significantly contributed to better management and treatment outcomes.

In recent years, the survival rate of multiple myeloma patients has notably improved with the introduction of novel agents for induction therapy and high-dose chemotherapy followed by ASCT (Turesson *et al.*, 2010; Kumar *et al.*, 2008; Gay *et al.*, 2015; Lehners *et al.*, 2018; Child *et al.*, 2003). ASCT plays a critical role in managing younger patients with multiple myeloma, leading to better progression-free survival (PFS) and overall survival (OS) compared to conventional chemotherapy (CC) (Mina *et al.*, 2015; Gupta *et al.*, 2009; Gay *et al.*, 2017).

Typically, when initiating treatment for multiple myeloma, patients undergo a four to six-month induction therapy using a combination of novel agents such as lenalidomide, thalidomide, bortezomib, and dexamethasone. For patients under 65 years old, high-dose melphalan treatment followed by ASCT is recommended after induction therapy. Additionally, maintenance therapy with lenalidomide/thalidomide, lenalidomide, or bortezomib is often given for 1-2 years after the initial treatment (Kumar *et al.*, 2008; Palumbo and Anderson, 2011; Christoph Röllig, 2015). Numerous randomized and non-randomized trials, population-based studies, and meta-analyses have provided evidence supporting the effectiveness of this treatment regimen, demonstrating high response rates, improved overall survival, and progression-free survival (Attal *et al.*, 1996; Child *et al.*, 2003; Fermand *et al.*, 2005). However, some patients experience relapse within two years of undergoing the graft (Huang *et al.*, 2017), necessitating an examination of key factors contributing to the higher risk of early relapse.

This study aims to identify the key clinical factors responsible for the early relapse of multiple myeloma patients who underwent ASCT . We analyzed the clinical data of 253 patients diagnosed with multiple myeloma, comprising 166 males and 87 females, with a median age of 52 years. These patients received ASCT treatment at the Department of Medical Oncology, All India Institute of Medical Sciences (AIIMS), between August 2005 and December 2016. To achieve our objective, we employed Boruta, Random Forest, and Bayesian network methods to assess the significance of features and identify the direct and indirect factors contributing to early relapse (Kursa and Rudnicki,

2010; Breiman, 2001; Friedman *et al.*, 1997).

Our analysis revealed that two factors, Primary and diabetes, were directly linked to an elevated risk of relapse. The Primary factor was categorized into two groups: patients who received stem cell transplantation after initial treatment and patients who received salvage therapy followed by stem cell transplantation after relapse. Additionally, we found that the Glomerular Filtration Rate (GFR) score, representing the change in GFR over induction therapy, was indirectly related to early relapse. These findings provide valuable insights for clinicians to identify patients at high risk of early relapse and develop effective treatment strategies to enhance patient well-being.

4.2 Materials and methods

4.2.1 Patients

From April 1990 to December 2016, the Department of Medical Oncology at All India Institute of Medical Sciences (AIIMS) treated 347 individuals diagnosed with MM through ASCT. All patients included in the study provided written consent. The Institute of Ethics Committee at All India Institute of Medical Sciences (AIIMS) approved the study with the reference number IEC-523/05.10.2018. The pre-transplant features of the patients are presented in **Table 4.1**.

4.2.2 Transplant protocol

All patients were initially assessed in the weekly Bone Marrow (BM) Transplant Clinic, where the medical staff informed patients and their family members about the potential advantages and drawbacks of bone marrow transplantation. Before the transplant, patients underwent a comprehensive pre-transplant evaluation that involved a detailed medical history, a thorough physical examination, and a determination of the disease stage based on both the International Staging System (ISS) (Greipp *et al.*, 2005) and the Durie and Salmon (DSS) (Durie and Salmon, 1975). The patient's previous treatment details were documented. During the pre-transplant evaluation, a range of tests was conducted, including tests for quantitative immunoglobulin levels, serum β-2 mi-

croglobulin, immune-fixation studies, serum and urine electrophoresis, a skeletal survey, bone marrow examination, renal and liver function tests, measurements of total and differential blood counts and hemoglobin levels. The patients provided written consent before these tests were conducted. Adverse effects related to the treatment were evaluated based on the Seattle criteria (Bearman *et al.*, 1988). The majority of patients received stem cells sourced from granulocyte colony-stimulating factor (G-CSF), while a small number of patients received stem cells from bone marrow. Peripheral blood stem cells mobilized by cyclophosphamide were used for stem cell harvesting in fewer than 10 patients. An even smaller number of patients had their stem cells harvested from bone marrow. The viability of cells was assessed using the trypan blue dye exclusion test (Kumar *et al.*, 2009). The initial treatment, which lasts for 4-5 months, generally involves 4-6 cycles. During this time, patients receive a combination of novel therapies such as proteasome inhibitors (bortezomib), immune modulators (thalidomide, lenalidomide), and dexamethasone. Following this, patients receive high doses of melphalan (Kumar *et al.*, 2016). The conditioning regimen prior to transplantation involved administering melphalan at a dosage of 150-225mg/m^2 to 218 patients (86.2%) through a slow intravenous push on the first day, followed by forced alkaline diuresis. Patients with renal insufficiency ($eGFR$<$40ml/min/1.73m^2$) at the time of transplantation received a lower dose of melphalan, specifically 120-150 mg/m^2 (Levine *et al.*, 1993). No significant difference was observed in the outcome of progression-free survival (PFS) and overall survival (OS) with the change in melphalan dosage, which is consistent with previous literature (Agrahari *et al.*, 2018). After the mobilization and collection of peripheral blood stem cells, they were then infused into the patient's bloodstream through an intravenous (i.v.) route, 24 hours after the administration of a high dose of melphalan. All patients received subcutaneous injections of G-CSF at a dose of 5 μg/kg daily, including on the day of stem cell infusion (day 0), 12 hours after the infusion, and continuously thereafter until engraftment. Isolation rooms were used to treat patients, and healthcare workers followed reverse barrier nursing practices. **Figure 4.1** illustrates a typical treatment plan for individuals diagnosed with multiple myeloma.

4.2.3 Data pre-processing

In order to determine the most significant factor affecting progression-free survival, we analyzed 40 variables extracted from clinical and laboratory data, as listed in **Table 4.1**. We confirmed that the data associated with these variables are usually obtainable during the pre-transplant phase. Nevertheless, some variables had missing values, which we handled by utilizing an R package that implements MICE, a widely adopted algorithm for imputing missing values (van Buuren and Groothuis-Oudshoorn, 2010). We transformed 29 categorical variables into numerical ones using one hot encoding (Seger, 2018). This step is crucial for enabling machine learning algorithms to operate effectively.

Table 4.1: Summary of the pretransplant clinical factors.

Feature	Description	Numeric range	Standard deviation
Age	Refers to the age of the patient.	29-68	8.43202
Albumin	Refers to the level of albumin in the blood during the patient's initial visit, where lower serum albumin levels are associated with clinical factors reflecting disease severity in multiple myeloma patients.	1.01-5.70	0.79976
Body Mass Index (BMI)	Refers to the BMI of the patient.	14.5-39.34	4.151437
Body surface area (BSA)	Refers to the body surface area of the patient.	1.17-11.61	0.5653937
GFR Grade	Refers to the categorized glomerular filtration rate level (5=<15ml, 4=15-29, 3=>=30 to 59, 2>=60 to 89, 1=>=90ml).	1-5	NA
BMPC	Refers to the count of bone marrow plasma cells while harvesting.	1-100	27.42393

Continued on next page

Table 4.1 – continued from previous page

Feature	Description	Numeric range	Standard deviation
Pre-transplant creatinine level	Refers to the creatinine level after induction therapy and before autologous stem cell transplantation.	0.49-6.10	0.4590685
Creatinine	Refers to the level of creatinine in the blood during the patient's initial visit, where high levels indicate kidney malfunctioning.	0.20-23.60	2.323097
Melphalan dose	Refers to the dose of melphalan given as the conditioning regimen after collecting stem cells.	100.00-225.0	20.91845
Symptom duration	Refers to the duration of symptoms in months.	0.0-30.0	5.65912
DSS	Refers to the Durie-Salmon Staging System, which calculates the staging of multiple myeloma patients by measuring blood calcium level, hemoglobin level, M protein level, and kidney function. Higher DSS stages are associated with higher severity (1=IA,2=IB,3=IIA,4=IIB,5=IIIA,6=IIIB,7=Not known).	1-7	NA
Gender	Refers to the gender of the patient (M= Male, F= Female).	M,F	NA

Continued on next page

Table 4.1 – continued from previous page

Feature	Description	Numeric range	Standard deviation
GFR Score	Refers to the categorical variable that shows how GFR changed over the induction therapy (1 <- GFR =40 ml base line and pre Transplant, 2 <- GFR<40 ml at base line, >40 ml pre Transplant, 3 <- GFR <40 ml base line and pre Transplant).	1-3	NA
Pre-transplant GFR	Refers to the glomerular filtration rate after induction therapy and before autologous stem cell transplantation.	5.0-187.84	26.67148
Glomerular Filtration Rate (GFR)	Refers to the glomerular filtration rate when the patient comes for the first time, which represents renal conditions where lower GFR levels are associated with clinical factors reflecting disease severity.	1.66-182	32.441
Height	Refers to the height of the patient.	140.0-190.0	9.967419
Hb	Refers to the hemoglobin level (g/dl) in the blood during the patient's initial visit.	3.20-16.0	2.601038
Line of induction therapy	The order in which induction therapy is given to a patient with cancer, with the first line being the initial treatment to reduce the tumor burden and prepare for further treatment.	1-3	NA

Continued on next page

Table 4.1 – continued from previous page

Feature	Description	Numeric range	Standard deviation
Line regimen	A specific combination of drugs and dosages used in a particular line of cancer treatment to achieve the best possible outcome with minimal side effects (1<- one line,2 <- Two lines,3 <- Three lines,4 <->3 lines)	1-4	NA
Response to induction therapy	Refers to how patients respond to induction therapy. The European Group for Blood and Bone Marrow Transplantation (EBMT) criteria were used to evaluate the response to ASCT six weeks after the transplant (Milpied *et al.*, 1996). Possible responses include 1=complete response (CR), 2=very good partial response (VGPR), 3 = partial response(PR), 4=stable disease, or 5 =no response (NR)	1-5	NA
Immunoglobulin type	Refers to the type of immunoglobulin isotype present in the blood(1 =IgG-Kappa, 2= IgG-Lambda,3=IgA-Kappa, 4=IgA-Lambda, 5=Kappa, 6= Lambda).	1-6	NA
ISS	The International Staging System measures serum albumin and beta-2-microglobulin levels to calculate three stages of multiple myeloma. Stage I is associated with less severity, while stage III is associated with the highest severity.	I-III	NA

Continued on next page

Table 4.1 – continued from previous page

Feature	Description	Numeric range	Standard deviation
Absolute lymphocyte count (ALC)	Refers to the lymphocyte count when a patient first comes for treatment.	230.0-21960.0	1839.455
Monoclonal protein level	Refers to the level of monoclonal protein in the blood when a patient first comes for treatment. Higher levels of monoclonal protein are associated with greater severity of the disease.	0.0-13.5	2.384272
Absolute Neutrophil Count (ANC)	Refers to the neutrophil count when a patient first comes for treatment.	0-37	4.224557
Platelet count	Refers to the number of platelets when a patient first comes for treatment.	13.0-824.0	107.518
CD34 cell number	Refers to the number of CD34 cells in stem cell harvesting. It indicates the purity of stem cells in the sample.	0.3018-16.7	2.107692
Primary	Patients are categorized into two groups: 1= those who received treatment and went on to receive stem cell transplant and 2= those who relapsed after initial treatment and then received salvage therapy followed by a stem cell transplant.	1-2	NA
Number of Cycles in 1st line treatment	Refers to the number of cycles of induction therapy a patient receives in the first line of treatment. Regimens given in one cycle are combinations of different drugs.	1-4	NA

Continued on next page

Table 4.1 – continued from previous page

Feature	Description	Numeric range	Standard deviation
Serum Calcium	Refers to the serum calcium level in the blood when a patient first comes for treatment. High calcium levels are related to advanced multiple myeloma and show osteoclast activity.	4.6-18.7	1.610736
Pre-transplant M-protein level	Refers to the serum M protein level after induction therapy and before autologous stem cell transplantation.	0.0-4.9	0.8888799
Stem cell Harvest site	Refers to the site from which stem cells are harvested, whether it is mobilized peripheral blood stem cells or bone marrow.	A-E	NA
Beta-2-microglobulin (B2M)	Refers to the level of beta-2-microglobulin in the blood when a patient first comes for treatment. B2M is the most powerful prognostic predictor of multiple myeloma, with lower serum albumin levels in multiple myeloma patients being associated with clinical factors reflecting disease severity (Greipp *et al.*, 2005)(Greipp et al. 2005).	900.0-325578.0	18164.87
Weight	Refers to the weight of the patient.	34.0-115.0	12.70282

Continued on next page

Table 4.1 – continued from previous page

Feature	Description	Numeric range	Standard deviation
Diabetic	Indicates whether the patient has diabetes or not. The numbers 0, 1, and 2 indicate different categories of diabetes status. 0 means the individual does not have diabetes, 1 means they have diabetes mellitus, and 2 means they have steroid-induced diabetes, which can be caused by long-term use of steroid medication.	0-2	NA
Hypertension	Indicates whether the patient has hypertension (1) or not(2).	1-2	NA
Dialysis	Indicates whether the patient receives dialysis=1 or not=2. Renal failure is the main reason for relapse in multiple myeloma patients (Dimopoulos *et al.*, 2010).	1-2	NA
Erythropoietin Treatment (Epo)	Indicates whether the patient received erythropoietin treatment=1 or not=2. Epo treatment is associated with improved immunological functions.	1-2	NA
Radiation Therapy	Indicates whether the patient received radiation therapy=1 or not=2. Radiation therapy is given in preparation for stem cell transplantation to kill myeloma cells.	1-2	NA
Extramedullary disease (EMD)	Indicates whether the patient has an extramedullary disease =1 or not =2.	1-2	NA

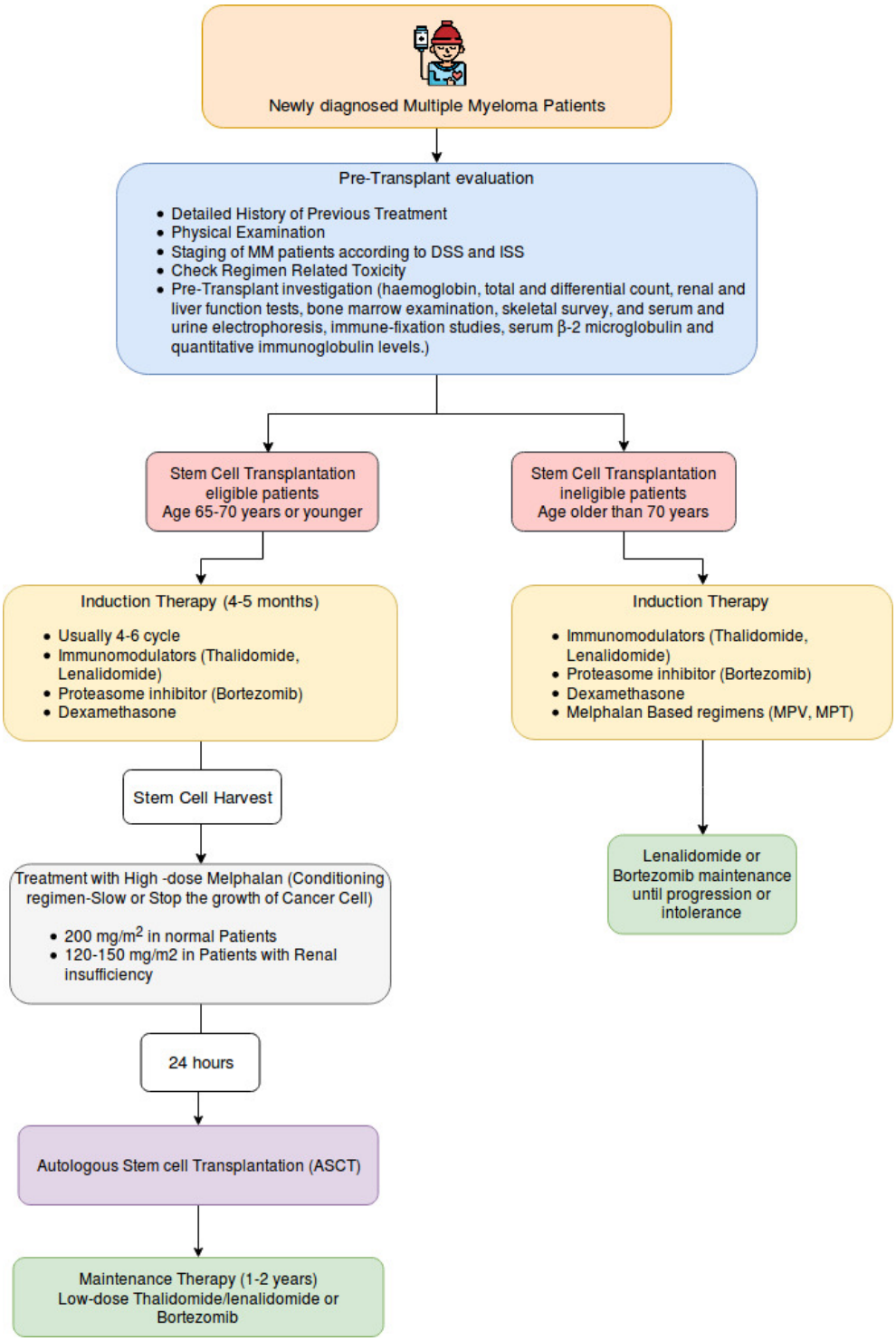

Figure 4.1: Multiple myeloma treatment approach. Standard autologous stem cell transplant procedure for individuals with multiple myeloma.

4.2.4 Univariate analysis

Our methodological approach aimed at evaluating the correlation of individual factors with observed Overall Survival (OS) and Progression-Free Survival (PFS) metrics. Employing the established Kaplan-Meier survival analysis technique (Goel *et al.*, 2010), we meticulously analyzed the associations between individual factors for OS and PFS. Categorical variables underwent thorough categorization, while for a subset of the dataset's numerical variables (23 out of 40), we utilized univariate *k-means* analysis (Wang and Song, 2011) to explore potential groupings. This analysis technique aimed to unveil inherent groupings or clusters within numeric data.

The process of binarisation in our analysis involved assigning two clusters for each numerical variable to simplify data interpretation. By setting the number of clusters to two for each variable, we determined a critical cut-off value based on the highest observed value within the cluster containing smaller data points. This value, referred to as 'C,' delineated two distinct ranges: data points less than or equal to 'C' were categorized as $\leq C$, while those surpassing 'C' were labeled as $> C$. Essentially, binarisation facilitated the simplification and transformation of complex numerical data into binary outcomes, enabling clearer categorization for subsequent analysis and interpretation.

4.2.5 Feature selection and Bayesian network analysis for PFS after autologous stem cell transplantation

We conducted a study to identify the most relevant predictors associated with progression-free survival (PFS) after ASCT in patients. We utilized pretransplant data from 253 patients with 40 clinical variables to classify patients into two groups: those who did not experience PFS relapse within 30 months (Class 0) and those who experienced PFS relapse (Class 1). We used the Boruta feature selection technique (Kursa and Rudnicki, 2010) with the response variable 'Class' set as a factor for binary classification to identify important predictors. The Boruta function in the Boruta package was run with 1000 maximum runs, and the most relevant predictors were obtained using the *getSelectedAttributes()* function.

To build a predictive model for progression-free survival in multiple myeloma pa-

tients, we used a subset of pretransplant data consisting of the most important predictors selected by Boruta, along with the PFS_Event variable. The Bayesian Network (BN) model (Friedman *et al.*, 1997) was constructed using the R package, bnlearn (Scutari, 2009). We fitted the BN model to the subset dataset using the *bn.fit()* function, which estimated the network structure parameters. To visualize the resulting model, we used *graphviz.plot()* function from the bnlearn package, with the highlight argument, applied to emphasize the 'PFS_Event' node and its ancestors. The BN model allowed us to identify the conditional dependencies between the selected attributes and the 'PFS_Event' variable, providing valuable insights into the factors that may influence progression-free survival after transplantation.

4.2.6 Random forest model for PFS prediction after stem cell transplantation using variable importance measures

We employed the random forest algorithm implemented in R programming language using the randomForest package (Liaw *et al.*, 2002) to construct a classification model for accurately predicting the progression-free survival (PFS) of multiple myeloma patients who have undergone autologous stem cell transplantation. Before training the model, we split the dataset into training and testing subsets in an 80:20 ratio. The binary 'Class' variable was defined to assign patients with PFS greater than or equal to 30 months to one class and others to the remaining class. The random forest model was trained on the training set while setting the *importance= TRUE* to compute the variable importance measures. Subsequently, we determined the importance of each feature using the Mean Decrease Gini index (Gastwirth, 1972) through the varImpPlot function from the varImp package (Probst, 2020). The top five most significant features were selected based on their Mean Decrease Gini index. A new random forest model was trained with only those five features and the 'Class' variable utilizing the *randomForest()* function from the randomForest'package (Liaw *et al.*, 2002). Finally, we evaluated the model's performance on the testing set using the confusion matrix and related metrics from the caret, R package (Kuhn, 2015).

4.3 Results

4.3.1 Patient characteristics

Autologous stem cell transplants have shown a consistent increase in number from 1990 to 2016, with a significant rise observed between 2010-2014, where the yearly count doubled compared to the previous decade. During the ASCT transplant protocol, patients were given a combination of drugs during induction therapy, with a current trend of utilizing novel agents. From April 1990 to March 2005, the VAD regimen, including Vincristine (V), Doxorubicin (A), and Peroral Dexamethasone (D), was administered to 72 patients, while alkylating agents were given to only 22 patients between July 1997 and March 2014. In contrast, 253 patients were treated with novel agents from August 2005 to December 2016, as presented in **Table 4.2**. Using novel agents has resulted in better survival rates than VAD and alkylating agents, as shown in **Figure 4.2**.

Table 4.2: The table describes the frequency of different combinations of novel agents in treating multiple myeloma.

Drug Combination	Agent Type	Number of Patients
Thalidomide/Dexamethasone	Two drug	180
Lenalidomide/Dexamethasone	Two drug	94
Bort/Dexa	Two drug	55
Bortezomib/Thalidomide/ Dexamethasone (VTD)	Three drug	31
Bortezomib/Lenalidomide/ Dexamethasone (VRD)	Three drug	71
Bortezomib/Cyclophosphamide/ Dexamethasone (VCD)	Three drug	23
Cyclophosphamide/Thalidomide/ Dexamethasone (CTD)	Three drug	23
Bortezomib/Low-dose dexamethasone/Pegylated liposomal doxorubicin (PAD)	Three drug	20
Melphalan/Prednisone/ Thalidomide (MPT)	Three drug	2
Bendamustine/Bortezomib/ Dexamethasone (BVD)	Three drug	1
Bortezomib/Thalidomide/ Cyclophosphamide/ Dexamethasone (VTCD)	Four drug	2

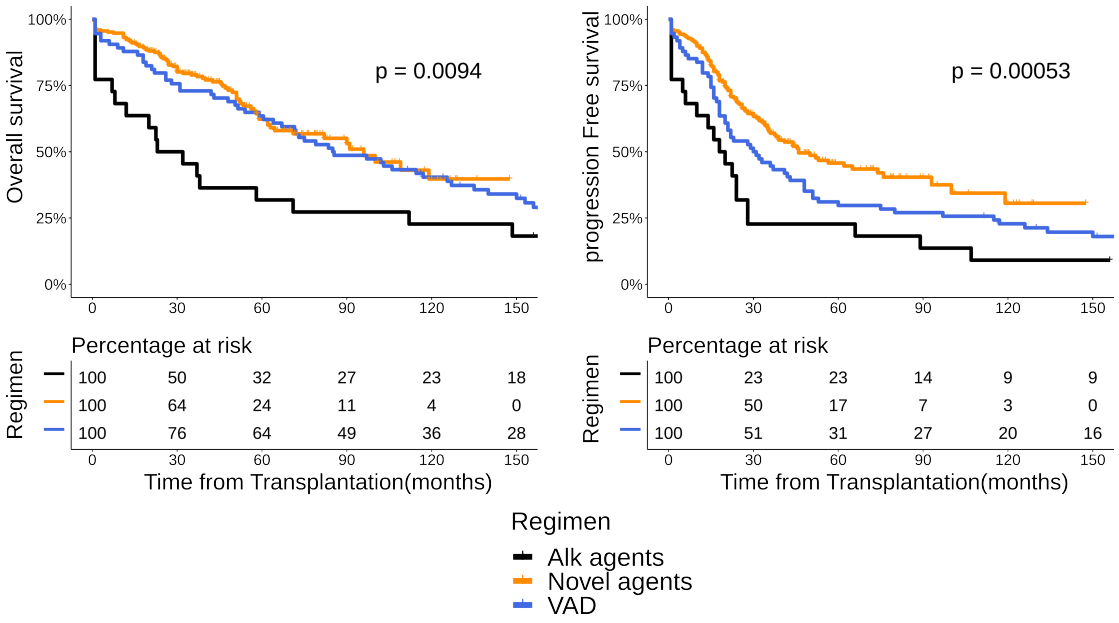

Figure 4.2: Survival Analysis Across Varied Drug Regimens in Patient Treatment. Comparison of overall survival (OS) and progression-free survival (PFS) among patients treated with alkylating agents, VAD, and novel agents.

Since induction therapy utilizing novel agents has become the most popular mode of treatment over the last decade, this study focuses on the cohort of 253 patients who underwent this treatment. There were no significant differences in survival trends among the novel agents, and no patients were lost during the follow-up period. The follow-up period was extended until November 30th, 2017, and **Table 4.3** highlights various critical patient characteristics.

4.3.2 Univariate analysis of prognostic factors impacting overall survival and progression-free Survival

To identify the factors that hold prognostic significance, we conducted a Kaplan-Meier survival analysis for each factor. Out of the 40 factors considered, 23 were numerical, while the remaining were categorical. We utilized univariate *k-means* to group patients based on each numerical feature and analyzed overall survival and progression-free survival for each factor. Certain factors demonstrated significant prognostic value for progression-free survival (PFS) and overall survival (OS). These included the number of pretransplant regimens used ($P < 0.001$ for both PFS and OS), pre-transplant M-protein level ($P = 0.0018$ for PFS and $P = 0.002$ for OS), the occurrence of relapse following

Table 4.3: Patient Characteristics

Characteristics	Number of Patients
Median Age	52
Gender	166 males, 87 females
Relapsed after remission, then transplant	69
Diabetic	107
Presence of Extramedullary Disease	57
Renal Condition (Patients required Dialysis)	8
ISS Stage during diagnosis	67 Stage-I, 94 Stage-II, 92 Stage-III
Line of Induction Therapy	174 one-line, 54 two-line, 19 three-line, 6 more than three line
Immunoglobulin Type	101 IgG-Kappa, 43 IgG-Lambda, 26 IgA-Kappa, 15 IgA-Lambda, 45 Kappa, 23 Lambda

remission labeled as primary ($P < 0.001$ for both OS and PFS) and serum albumin level ($P < 0.001$ for both OS and PFS). It is interesting to note that the response to induction therapy displayed a greater predictive value for PFS (with a P-value of 0.0015) compared to OS (with a P-value of 0.012). ISS staging appeared to have minimal predictive value for PFS. The key results of the univariate analyses have been summarized in **Table 4.4**. According to our analysis, we found that the *k-means* based grouping approach was scientifically valid and consistent with previous research findings. As an example, we identified a serum albumin level cut-off of 3.5 g/dL using *k-means*-based grouping, which is consistent with a prior study reporting that levels of serum albumin $\leq$3.5 g/dL were associated with increased mortality (Kim *et al.*, 2010).

4.3.3 Key risk factors for early relapse in multiple myeloma patients undergoing autologous stem cell transplantation

Our study focused on identifying crucial risk factors associated with relapse of autologous stem cell transplantation treatment in multiple myeloma patients. To achieve this, we divided the data into two classes: Class 1 included ASCT-treated patients who did not experience a relapse within 30 months of transplantation, while Class 2 included those who did relapse within the same time frame. We utilized two different methods, the Boruta R package (Kursa and Rudnicki, 2010) and the Random Forest algorithm

Table 4.4: Displaying univariate survival analysis for overall survival and progression-free survival used to evaluate the prognostic ability of individual factors.

Factors		n	Progression Free Survival					Overall Survival				
			Events	Median	p-value			Events	Median	p-value		
Primary	1st line transplantation	184	74	76	4.5e-11			43	135	2.2e-12		
	Relapse then transplantation	69	46	42				39	37			
Line Regimen	1 type	174	69	76	1.3e-06			41	135	9.3e-08		
	> 1 type	79	51	24				41	47			
Albumin	<3.5	103	58	34	0.00033			45	56	6.9e-06		
	>=3.5	150	62	76				37	119			
Response Induction	CR + VGPR	142	57	91	0.0015			36	100	0.012		
	Others	111	63	35				46	63			
Pretransplant M Spike	<0.25 g/dL	175	73	56.5	0.0018			43	>90	0.002		
	>=0.25 g/dL	78	47	28.0				39	62			
ISS	ISS-I	67	29	76		ISS-I	ISS-II	16	130		ISS-I	ISS-II
	ISS-II	94	48	44	ISS-II	0.2	-	34	91.5	ISS-II	0.0556	-
	ISS-III	92	43	36	ISS-III	0.15	0.42	32	62	ISS-III	0.0042	0.1129
EMD	Present	57	32	35.5	0.042			25	57	0.014		
	Absent	196	88	53				57	100			
Hb (g/dl)	<8.4	82	44	34	0.046			30	96	0.029		
	>=8.4	171	76	62				52	100			
Serum M Spike	<=1.91 g/dL	95	37	76	0.077			22	>100	0.071		
	>1.91 g/dL	158	83	39				60	90			
Beta 2 Microglobulin (mcg/L)	<4.2 mcg/mL	129	59	62	0.11			37	135	0.0095		
	>=4.2 mcg/mL	124	61	44				45	64			
GFR (mL/min/1.73 m^2)	<60	158	72	56.5	0.21			47	96	0.06		
	>=60	95	48	44				35	64.5			
Immunoglobulin Type	Kappa	172	79	53	0.41			35	64.5	0.076		
	Lambda	81	41	46				31	63			

(Liaw *et al.*, 2002), to identify the most important features. Boruta is a powerful feature selection algorithm that is specifically designed to handle noisy and redundant datasets. On the other hand, the Random Forest algorithm is based on the Gini impurity index, which measures the degree of node impurity in a decision tree. It is noteworthy that some of the features selected by the Gini impurity index, namely Primary, Diabetic, Pre_Tx_GFR, and ALC_baseline, are among the seven most significant attributes obtained from the Boruta algorithm **(Table 4.1, Figure 4.3)**. These 7 attributes are GFR_Score, Primary, Diabetic, Line_Regimen, Serum_Calcium, Pre_Tx_Mspike, and ALC_baseline **(Table 4.1, Figure 4.4)**. Among these attributes, Primary and Diabetic were found to be particularly important. However, Albumin was another top attribute determined by the Random Forest.

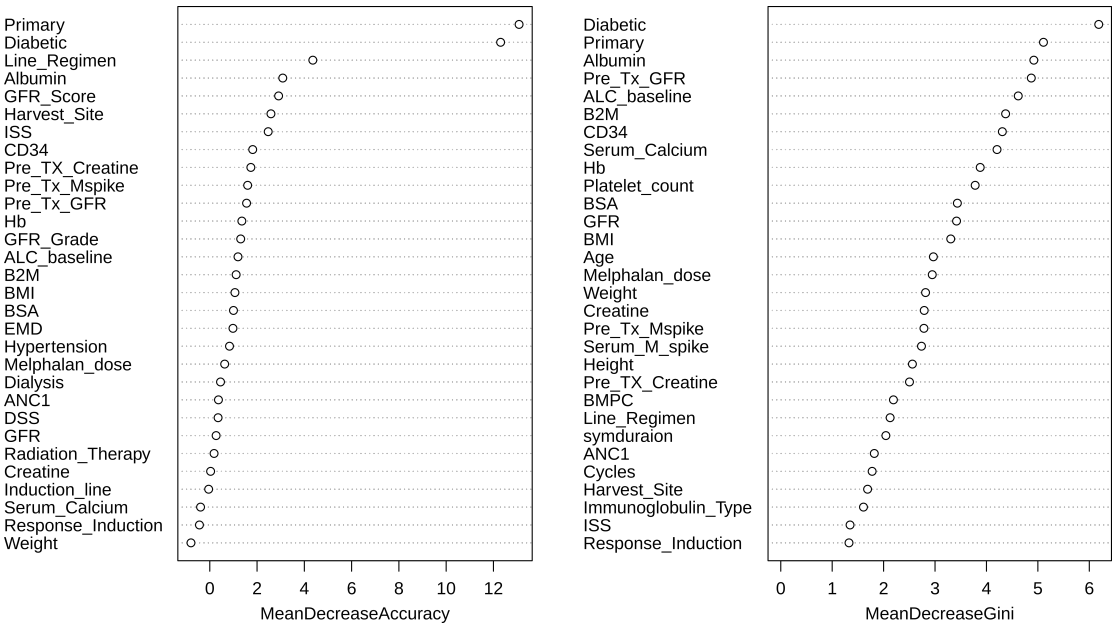

Figure 4.3: Predicting ASCT relapse: Variable importance plot Variable Importance Plot showing the top predictors of ASCT relapse identified by Random Forest analysis.

Our results were consistent with previous studies (Knudsen *et al.*, 2000; Wu *et al.*, 2014; Kropff *et al.*, 2003; Zagouri *et al.*, 2017; Coresh *et al.*, 2014). Studies have found that kidney dysfunction is a frequent complication of multiple myeloma, and a decline in glomerular filtration rate (GFR) greater than 25% is significantly associated with shorter progression-free survival (PFS) in multiple myeloma patients (Dimopoulos *et al.*, 2010). The study suggests that changes in GFR may predict disease progression, as confirmed by other studies (Schmidts *et al.*, 2019; Goswami *et al.*, 2019; DeFronzo *et al.*, 1978). Moreover, GFR is not only a prognostic factor but also a determinant in se-

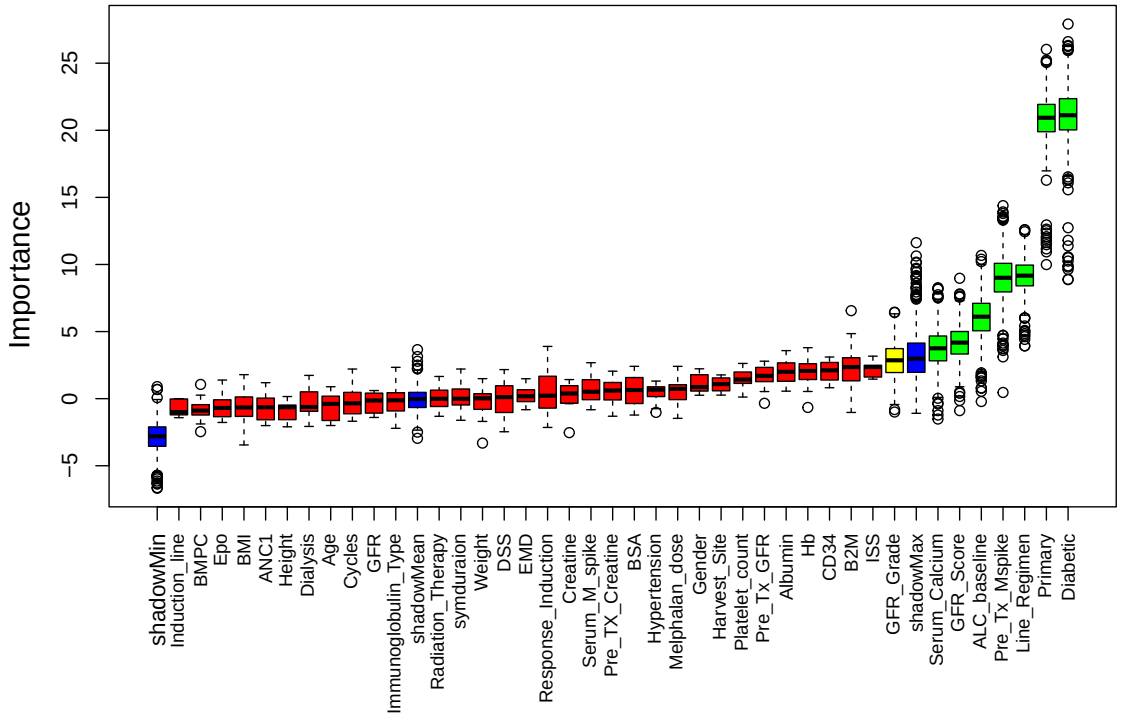

Figure 4.4: Feature significance: Boruta analysis for ASCT relapse. This boxplot illustrates the significance attributed to each feature. Columns highlighted in green are designated as 'confirmed,' while those in red are not deemed significant. Notably, a few blue bars, namely ShadowMax and ShadowMin, although not actual features are employed by the Boruta algorithm to determine the importance of a variable. The Boruta feature selection method has identified seven pivotal features associated with relapse following ASCT treatment in patients with multiple myeloma

lecting treatment for multiple myeloma patients (Katagiri *et al.*, 2013; Gonsalves *et al.*, 2015). In patients with advanced multiple myeloma, hypercalcemia or high calcium levels in the blood, is a common observation (Oyajobi, 2007; Mirrakhimov, 2015). This is due to excess secretion of parathyroid hormone-related protein (PTHrP) by cancerous plasma cells, which causes calcium to be released from the bones into the bloodstream (Kitazawa *et al.*, 2002). Hypercalcemia is linked to a poor prognosis and increased mortality in multiple myeloma patients (Bao *et al.*, 2020). Additionally, hypercalcemia is associated with bone lesions and reduced kidney function, which are important factors to consider in the context of ASCT treatment relapse (Dellay and Groth, 2016; Zagouri *et al.*, 2017). Line regimen refers to the specific combination of drugs and dosages used in a particular line of cancer treatment to achieve the best possible outcome with minimal side effects. Categorizing patients consistently based on their past treatments is crucial in clinical trials but can be challenging due to the various drugs and combinations of drugs used. To compare different treatments, a new categorization based on the drug classes the patients have been exposed to may be needed, which requires analyzing large amounts of data (Kumar *et al.*, 2022). Multiple myeloma patients with lower serum albumin levels have been shown to have more advanced disease, a more aggressive course, and a poorer response to treatment (Kim *et al.*, 2010; Solmaz *et al.*, 2023). Low albumin levels may indicate a compromised immune system, inflammation, and malnutrition, which can contribute to tumor growth and reduce the efficacy of cancer treatment (Wiedermann, 2021; Chen and Magalhaes, 1990). Studies have suggested that ALC baseline and pre-transplant M-spike levels may serve as prognostic markers for MM patients at a higher risk of relapse after ASCT (Dosani *et al.*, 2017; Rodon *et al.*, 2001). A low ALC baseline is associated with an increased risk of ASCT relapse, while a higher level of pre-transplant M spike is associated with an increased risk of relapse after ASCT (Rodon *et al.*, 2001). These markers could be useful for identifying high-risk MM patients who may require more intensive monitoring and treatment.

Primary and Diabetic are the two pivotal variables that significantly impact the risk of progression-free survival in multiple myeloma patients. Primary classifies patients into two groups based on their treatment history, while Diabetic reveals the presence or absence of diabetes **(Table 4.1)**. Diabetes mellitus and steroid-induced diabetes can negatively impact multiple myeloma treatments in various ways, including increased infection risk, immune system impairment, drug interactions, and the need for addi-

tional monitoring. Healthcare providers must consider diabetes status when creating a treatment plan for multiple myeloma patients. Patients with multiple myeloma can be classified into two groups based on their treatment history: those who initially received treatment followed by ASCT and those who received salvage therapy before undergoing ASCT. ASCT is commonly used as consolidation therapy to prevent relapse after initial treatment (Mohty and Harousseau, 2014). Patients who relapsed after initial treatment and received salvage therapy before ASCT may have a higher risk of relapse due to disease resistance and burden (Farris *et al.*, 2019; Rajkumar and Kumar, 2020).

4.3.4 Partially mediated role of glomerular filtration rate in the association between diabetes and early relapse

In order to predict progression-free survival, we utilized Bayesian networks (Friedman *et al.*, 1997) to explore the relationships between the seven most crucial variables selected through Boruta, R package (Kursa and Rudnicki, 2010). The variable indicating whether relapse occurred in progression-free survival, known as a PFS event, was also incorporated into the analysis.

Through Bayesian network analysis, we determine the relationship between variables and their impact on progression-free survival (PFS) in multiple myeloma patients. Our analysis identified Primary and Diabetic as two pivotal variables that significantly impact PFS **(Table 4.1)**. Furthermore, the glomerular filtration rate score (GFR changed over the induction therapy), a measure of kidney function, was also identified as an important variable that directly affects PFS (Dimopoulos *et al.*, 2010) and indirectly worsens diabetes **(Figure 4.5)**.

Multiple myeloma is a cancer that can lead to renal impairment, a common complication of the disease. When kidney function is impaired, it can disrupt glucose regulation, leading to an increased risk of diabetes. Similarly, diabetes can exacerbate kidney damage and further compromise renal function, creating a vicious cycle (Alicic *et al.*, 2017; Tuttle *et al.*, 2014; Thomas *et al.*, 2015).

Our results align with previous studies that have found a link between diabetes, renal impairment, and an increased risk of relapse in multiple myeloma treatment. For instance, Issa et al. (Issa *et al.*, 2011) reported that patients with diabetes had a higher

incidence of early relapse following ASCT than non-diabetic patients. They suggested that high blood sugar levels in diabetic patients could increase the risk of infections and impair the immune system, impacting the effectiveness of chemotherapy and other treatments. Additionally, some diabetes medications and steroids can interact with multiple myeloma drugs, affecting their metabolism and side effects (Wu *et al.*, 2014).

Furthermore, kidney failure is a common complication of multiple myeloma and can be caused by the accumulation of abnormal proteins produced by cancer cells in the kidneys (Dimopoulos *et al.*, 2008). Patients with kidney failure may require more intensive treatment and monitoring and may be at a higher risk of relapse due to the more advanced stage of their disease (Dimopoulos *et al.*, 2010). Additionally, kidney failure can affect the metabolism and elimination of chemotherapy drugs, leading to higher toxicity levels and potential treatment delays (Hutchison *et al.*, 2011). Our analysis underscores the importance of monitoring kidney function and blood sugar levels in multiple myeloma patients with diabetes to prevent complications and improve overall health outcomes. By identifying Primary, Diabetic, and GFR as pivotal variables impacting PFS, our Bayesian network analysis provides valuable insights that can guide clinical decision-making and inform personalized treatment plans for multiple myeloma patients.

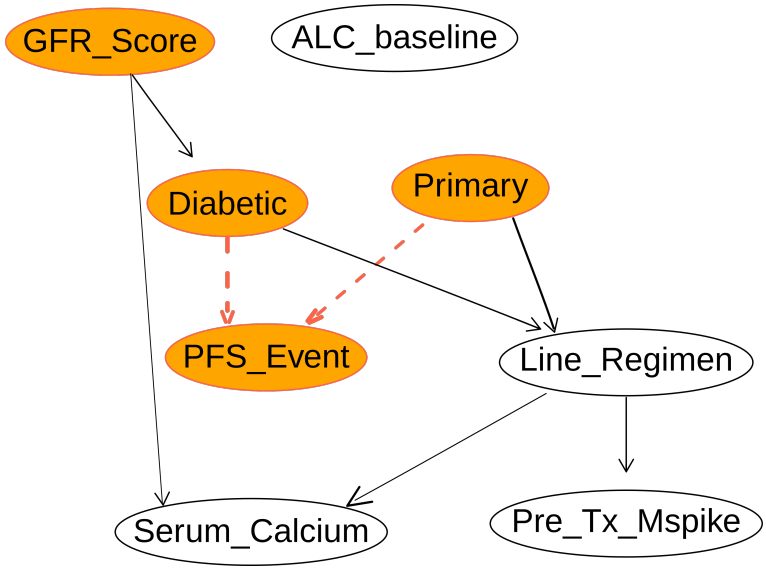

Figure 4.5: Analyzing PFS in multiple myeloma: Bayesian insights. Bayesian network analysis reveals the conditional relationships between significant clinical factors that directly or indirectly impact the risk of progression-free survival in multiple myeloma patients.

4.3.5 Random forest classification model

A Random forest classification model was used to select the most important features from training data and then subset the data to model with the selected variables **(Figure 4.3)**. The resulting model was used to predict outcomes on a test dataset, which was also subsetted only to include the five most important variables. The performance of the resulting binary classification model was evaluated using a confusion matrix and various statistical measures, including accuracy, 95% confidence interval, kappa, sensitivity, specificity, positive predictive value, negative predictive value, prevalence, detection rate, detection prevalence, and balanced accuracy. The model achieved an accuracy of 76.47%, with a 95% confidence interval and kappa of 0.5357. The sensitivity was 67.86%, specificity was 86.96%, and balanced accuracy was 77.41%. These results emphasize the significance of the five crucial variables identified by the Random forest analysis in predicting the outcome of the test dataset **(Figure 4.3)**. By incorporating these variables into predictive models, clinicians can enhance patient results and develop personalized treatment plans based on individual risk factors.

4.4 Discussion

The study aimed to identify the key risk factors associated with multiple myeloma. We used the Boruta R package (Kursa and Rudnicki, 2010) and the Random Forest algorithm (Liaw *et al.*, 2002) to identify the most important features. The top seven attributes determined by these algorithms were GFR_Score, Primary, Diabetic, Line_Regimen, Serum_Calcium, Pre_Tx_Mspike, and ALC_baseline **(Table 4.1)**. The results were consistent with previous studies that suggest that a decline in GFR greater than 25%, hypercalcemia, low serum albumin, and pre-transplant M spike levels may serve as prognostic markers for multiple myeloma patients (Knudsen *et al.*, 2000; Wu *et al.*, 2014; Kropff *et al.*, 2003; Zagouri *et al.*, 2017; Coresh *et al.*, 2014).

The results of our analysis indicate that Primary and Diabetic are the two most important variables affecting the risk of progression-free survival in multiple myeloma patients. Notably, Primary was found to be the most effective prognostic factor and showed a significant impact on univariate survival analysis. These findings suggest that MM patients who have experienced a relapse in previous treatment and then receive

salvage therapy may be at higher risk of treatment relapse. Therefore, it is crucial to take into account the patient's Primary and diabetes status when creating a treatment plan. Diabetes can have various negative impacts on multiple myeloma treatment, further highlighting the need for careful consideration of this comorbidity in treatment planning.

The study highlights the potential impact of kidney disease and diabetes on multiple myeloma treatment outcomes. Kidney failure, a common complication of multiple myeloma, may require more intensive treatment and monitoring and could lead to a higher risk of relapse due to the more advanced stage of the disease. Additionally, impaired kidney function can disrupt glucose regulation, leading to an increased risk of diabetes. Furthermore, diabetes can exacerbate kidney damage and further compromise renal function, creating a vicious cycle. In addition, diabetes can negatively impact multiple myeloma treatment in various ways, including increasing the risk of infections, impairing the immune system, and interacting with multiple myeloma drugs. This highlights the need for careful consideration of diabetes and kidney disease status when developing treatment plans for multiple myeloma patients.

CHAPTER 5

Conclusion

The primary objective of this thesis is to explore various machine-learning approaches for detecting and managing blood-based cancers. This includes exploring different types of algorithms, such as supervised and unsupervised learning, and evaluating their performance on real-world datasets. The thesis also aims to investigate the potential use of machine-learning techniques in personalized medicine, where individualized treatment plans can be developed based on a patient's specific genetic and clinical data.

Overall, the goal of this thesis is to contribute to the development of more effective and efficient methods for detecting and managing blood-based cancers, which can ultimately improve patient outcomes and quality of life.

5.1 Summary of contribution

The thesis at hand is organized into several chapters that explore different aspects of using machine-learning approaches to detect and manage blood-based cancers. In this section, we provide a brief summary of each chapter to provide a comprehensive view of the thesis as a whole.

5.1.1 Unsupervised clustering of circulating tumor cells from a large pool of blood cells

The detection and characterization of CTCs has become a crucial aspect in the treatment of cancer, as CTCs represent the only live cells that can provide information about the entire transcriptome and the cells that leave the bloodstream, leading to metastasis. Through the analysis of CTCs, we can monitor the progression of the disease, assess the response to treatment, and understand the spatiotemporal heterogeneity of tumor cells. Therefore, the ability to detect and isolate CTCs is a critical step in understanding the molecular mechanisms that govern cancer biology and designing effective

treatment strategies. The traditional methods for identifying CTCs using single-cell RNA sequencing have significant limitations due to molecular diversity, low CTC concentration, and down-regulation of epithelial markers. In response, a marker-agnostic approach was developed using deep dictionary clustering to detect CTCs in an unsupervised manner from a pool of white blood cells. This method transforms normalized and log-transformed expression vectors linked to CTCs into a pathway enrichment score vector, which is then used to cluster CTCs and white blood cells into distinct clusters. The method is robust, neutralizes batch effects, and offers a functional/ mechanistic perspective on unraveling cellular heterogeneity. The study compares unCTC to four existing integrative single-cell analytic pipelines, and unCTC outperformed them for the evaluated metrics. Using unCTC in CTC analysis is a promising development with the chance to improve our knowledge of cancer biology and ultimately elevate patient well-being.

5.1.2 Employing unCTC for marker-free characterization of circulating tumor cells

This chapter presents a validation of the unCTC pipeline, which uses a dataset obtained from the ClearCell® FX and Polaris™ systems to capture CTCs based on their size and exclude immune cells. CTCs from breast patients and WBCs from healthy patients were used for comparison, along with CTCs obtained using the Clear Cell Polaris workflow. The unCTC pipeline was found to do a great job of separating CTCs and WBCs into separate clusters with TPM and Count data, compared to well-known methods such as Seurat, fastmnn, and harmony for integrating scRNA-seq data. In addition, this chapter also used supervised and semi-supervised methods to further characterize the CTCs. The lineage validation was performed using several techniques such as single markers, sets of markers, differential genes, differential pathways, and copy number variation analysis. These techniques helped to validate the lineage of CTCs and distinguish them from immune cells. Based on the results, the chapter concludes that the unCTC pipeline is a reliable unsupervised method for clustering CTCs and identifying immune cells in different clusters, making it an excellent tool for further research and analysis in this field.

5.1.3 Identifying pre-transplant risk factors associated with early relapse in multiple myeloma patients undergoing autologous stem cell transplantation

The primary objective of cancer therapies is to eliminate or control cancer cells in the patient's body. While there may be various clinical factors that can impact treatment outcomes. In this chapter, the importance of considering clinical factors alongside cancer cells in the treatment of cancer is emphasized. It is worth noting that the treatment of cancer also involves various clinical factors that can worsen or even fail treatment, and managing these factors could lead to better outcomes and increased survival rates for patients. To investigate this further, the chapter employs machine-learning and statistical tools to identify clinical factors that are responsible for the relapse of ASCT therapy. The findings indicate that Glomerular Filtration Rate (GFR), diabetic condition, and relapse in earlier treatment for multiple myeloma can lead to early relapse of ASCT treatment compared to their opposite conditions. Moreover, the study suggests that managing additional factors such as albumin level, M protein level, line regimen (specific combination of drugs), serum calcium level, and ALC baseline can improve the success of ASCT treatment. Overall, the chapter highlights the importance of managing clinical factors in addition to cancer cells to enhance treatment efficacy and patient outcomes.

5.2 Future work

My research interests lie in further exploring the potential of machine learning algorithms in the context of pathways-based single-cell multi-omics studies to find spatial and temporal heterogeneity of cancer cells. With the advent of single-cell technologies, we have been able to obtain high-resolution molecular profiles of individual cells, providing a wealth of information about cellular heterogeneity within tumors. However, the challenge lies in integrating this multi-dimensional data to extract meaningful biological insights. Recent studies have shown that machine learning can play a crucial role in identifying patterns within these complex datasets and elucidating the molecular mechanisms that drive tumor progression and treatment resistance.

Another area of interest for my future work is the application of machine learning in spatial transcriptomics to understand the tumor microenvironment and the interaction of tumor cells with surrounding immune cells. The tumor microenvironment is a complex and dynamic ecosystem, composed of a variety of cell types, extracellular matrix components, and signaling molecules. Understanding the spatial distribution of these components is critical for developing effective therapeutic strategies. Spatial transcriptomics is an emerging technology that enables the simultaneous profiling of gene expression and spatial location of individual cells within a tissue section. Machine learning algorithms can be used to analyze these spatial transcriptomic datasets and identify cellular interactions, gene expression patterns, and spatial relationships that are relevant to tumor biology. Overall, machine learning will play an increasingly important role in advancing our understanding of cancer biology and ultimately lead to the development of more effective cancer therapies.

REFERENCES

1. **Abreu, M.**, **P. Cabezas-Sainz**, **L. Alonso-Alconada**, **A. Ferreirós**, **P. Mondelo-Macía**, **R. M. Lago-Lestón**, **A. Abalo**, **E. Díaz**, **S. Palacios-Zambrano**, **A. Rojo-Sebastian**, **R. López-López**, **L. Sánchez**, **G. Moreno-Bueno**, and **L. Muinelo-Romay** (2020). Circulating tumor cells characterization revealed TIMP1 as a potential therapeutic target in ovarian cancer. *Cells*, **9**(5).

2. **Aceto, N.** (2020). Bring along your friends: Homotypic and heterotypic circulating tumor cell clustering to accelerate metastasis. *Biomed. J.*, **43**(1), 18–23.

3. **Aceto, N.**, **A. Bardia**, **D. T. Miyamoto**, **M. C. Donaldson**, **B. S. Wittner**, **J. A. Spencer**, **M. Yu**, **A. Pely**, **A. Engstrom**, **H. Zhu**, **B. W. Brannigan**, **R. Kapur**, **S. L. Stott**, **T. Shioda**, **S. Ramaswamy**, **D. T. Ting**, **C. P. Lin**, **M. Toner**, **D. A. Haber**, and **S. Maheswaran** (2014). Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell*, **158**(5), 1110–1122.

4. **Adams, S. A.** and **C. Petersen** (2016). Precision medicine: opportunities, possibilities, and challenges for patients and providers. *J. Am. Med. Inform. Assoc.*, **23**(4), 787–790.

5. **Agrahari, R.**, **A. Foroushani**, **T. R. Docking**, **L. Chang**, **G. Duns**, **M. Hudoba**, **A. Karsan**, and **H. Zare** (2018). Applications of bayesian network models in predicting types of hematological malignancies. *Sci. Rep.*, **8**(1), 6951.

6. **Al-Ostoot, F. H.**, **S. Salah**, **H. A. Khamees**, and **S. A. Khanum** (2021). Tumor angiogenesis: Current challenges and therapeutic opportunities. *Cancer Treat Res Commun*, **28**, 100422.

7. **Alicic, R. Z.**, **M. T. Rooney**, and **K. R. Tuttle** (2017). Diabetic kidney disease: Challenges, progress, and possibilities. *Clin. J. Am. Soc. Nephrol.*, **12**(12), 2032–2045.

8. **Alimirzaie, S.**, **M. Bagherzadeh**, and **M. R. Akbari** (2019). Liquid biopsy in breast cancer: A comprehensive review. *Clin. Genet.*, **95**(6), 643–660.

9. **Alix-Panabières, C.** and **K. Pantel** (2013). Circulating tumor cells: liquid biopsy of cancer. *Clin. Chem.*, **59**(1), 110–118.

10. **Alix-Panabières, C.** and **K. Pantel** (2014). Technologies for detection of circulating tumor cells: facts and vision. *Lab Chip*, **14**(1), 57–62.

11. **Andrews, S.** (2010). Babraham bioinformatics-FastQC a quality control tool for high throughput sequence data. *URL: https://www. bioinformatics. babraham. ac. uk/projects/fastqc*.

12. **Arora, P.**, **D. Boyne**, **J. J. Slater**, **A. Gupta**, **D. R. Brenner**, and **M. J. Druzdzel** (2019). Bayesian networks for risk prediction using Real-World data: A tool for precision medicine. *Value Health*, **22**(4), 439–445.

13. **Atasoy, H.**, **B. N. Greenwood**, and **J. S. McCullough** (2019). The digitization of patient care: a review of the effects of electronic health records on health care quality and utilization. *Annual review of public health*, **40**, 487–500.

14. **Attal, M.**, **J.-L. Harousseau**, **A.-M. Stoppa**, **J.-J. Sotto**, **J.-G. Fuzibet**, **J.-F. Rossi**, **P. Casassus**, **H. Maisonneuve**, **T. Facon**, **N. Ifrah**, and **Others** (1996). A prospective, randomized trial of autologous bone marrow transplantation and chemotherapy in multiple myeloma. *N. Engl. J. Med.*, **335**(2), 91–97.

15. **Azuaje, F.** (2019). Artificial intelligence for precision oncology: beyond patient stratification. *NPJ Precis Oncol*, **3**, 6.

16. **Balic, M.**, **N. Dandachi**, **G. Hofmann**, **H. Samonigg**, **H. Loibner**, **A. Obwaller**, **A. van der Kooi**, **A. G. J. Tibbe**, **G. V. Doyle**, **L. W. M. M. Terstappen**, and **T. Bauernhofer** (2005). Comparison of two methods for enumerating circulating tumor cells in carcinoma patients. *Cytometry B Clin. Cytom.*, **68**(1), 25–30.

17. **Bankó, P.**, **S. Y. Lee**, **V. Nagygyörgy**, **M. Zrínyi**, **C. H. Chae**, **D. H. Cho**, and **A. Telekes** (2019). Technologies for circulating tumor cell separation from whole blood. *J. Hematol. Oncol.*, **12**(1), 1–20.

18. **Bao, L.**, **Y. Wang**, **M. Lu**, **B. Chu**, **L. Shi**, **S. Gao**, **L. Fang**, and **Q. Xiang** (2020). Hypercalcemia caused by humoral effects and bone damage indicate poor outcomes in newly diagnosed multiple myeloma patients. *Cancer Med.*, **9**(23), 8962–8969.

19. **Barrios, D.** and **C. Prieto** (2017). D3GB: An interactive genome browser for r, python, and WordPress. *J. Comput. Biol.*, **24**(5), 447–449.

20. **Bauckhage, C.** (2015). k-means clustering is matrix factorization. *arXiv preprint arXiv:1512.07548*.

21. **Baylin, S. B.** and **P. A. Jones** (2016). Epigenetic determinants of cancer. *Cold Spring Harb. Perspect. Biol.*, **8**(9).

22. **Bearman, S. I.**, **F. R. Appelbaum**, **C. D. Buckner**, **F. B. Petersen**, **L. D. Fisher**, **R. A. Clift**, and **E. D. Thomas** (1988). Regimen-related toxicity in patients undergoing bone marrow transplantation. *J. Clin. Oncol.*, **6**(10), 1562–1568.

23. **Bethesda, M. D.** (2018). SEER lifetime risk (percent) of being diagnosed with cancer by site and Race/Ethnicity: Both sexes, 18 SEER areas, 2012-2014 (table 1.15) national cancer institute.

24. **Bhosale, P. G.**, **S. Cristea**, **S. Ambatipudi**, **R. S. Desai**, **R. Kumar**, **A. Patil**, **S. Kane**, **A. M. Borges**, **A. A. Schäffer**, **N. Beerenwinkel**, and **M. B. Mahimkar** (2017). Chromosomal alterations and gene expression changes associated with the progression of leukoplakia to advanced gingivobuccal cancer. *Transl. Oncol.*, **10**(3), 396–409.

25. **Bi, W. L.**, **A. Hosny**, **M. B. Schabath**, **M. L. Giger**, **N. J. Birkbak**, **A. Mehrtash**, **T. Allison**, **O. Arnaout**, **C. Abbosh**, **I. F. Dunn**, *et al.* (2019). Artificial intelligence in cancer imaging: clinical challenges and applications. *CA: a cancer journal for clinicians*, **69**(2), 127–157.

26. **Bièche, I.**, **M. H. Champème**, and **R. Lidereau** (1995). Loss and gain of distinct regions of chromosome 1q in primary breast cancer. *Clin. Cancer Res.*, **1**(1), 123–127.

27. **Bittner, K. R.**, **J. M. Jiménez**, and **S. R. Peyton** (2020). Vascularized biomaterials to study cancer metastasis. *Adv. Healthc. Mater.*, **9**(8), e1901459.

28. **Bodard, S.**, **Y. Liu**, **S. Guinebert**, **Y. Kherabi**, and **T. Asselah** (2023). Performance of radiomics in microvascular invasion risk stratification and prognostic assessment in hepatocellular carcinoma: A meta-analysis. *Cancers*, **15**(3), 743.

29. **Bollschweiler, E.** (2003). Benefits and limitations of Kaplan–Meier calculations of survival chance in cancer surgery. *Langenbecks. Arch. Surg.*, **388**(4), 239–244.

30. **Bork, U.**, **N. N. Rahbari**, **S. Schölch**, **C. Reissfelder**, **C. Kahlert**, **M. W. Büchler**, **J. Weitz**, and **M. Koch** (2015). Circulating tumour cells and outcome in non-metastatic colorectal cancer: a prospective study. *Br. J. Cancer*, **112**(8), 1306–1313.

31. **Bottaci, L.**, **P. J. Drew**, **J. E. Hartley**, **M. B. Hadfield**, **R. Farouk**, **P. W. Lee**, **I. M. Macintyre**, **G. S. Duthie**, and **J. R. Monson** (1997). Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. *Lancet*, **350**(9076), 469–472.

32. **Breiman, L.** (2001). Random forests. *Mach. Learn.*, **45**(1), 5–32.

33. **Bulfoni, M.**, **M. Turetta**, **F. Del Ben**, **C. Di Loreto**, **A. P. Beltrami**, and **D. Cesselli** (2016). Dissecting the heterogeneity of circulating tumor cells in metastatic breast cancer: Going far beyond the needle in the haystack. *Int. J. Mol. Sci.*, **17**(10).

34. **Butler, A.**, **P. Hoffman**, **P. Smibert**, **E. Papalexi**, and **R. Satija** (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**(5), 411–420.

35. **Büttner, M.**, **Z. Miao**, **F. A. Wolf**, **S. A. Teichmann**, and **F. J. Theis** (2019). A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods*, **16**(1), 43–49.

36. **Cancer.org** (2018). Cancer facts and figures 2018. *American Cancer Society*.

37. **Caplan, R. J.**, **T. F. Pajak**, and **J. D. Cox** (1994). Analysis of the probability and risk of cause-specific failure. *Int. J. Radiat. Oncol. Biol. Phys.*, **29**(5), 1183–1186.

38. **Cardona, A. F.**, **O. Arrieta**, **A. Ruiz-Patiño**, **C. Sotelo**, **N. Zamudio-Molano**, **Z. L. Zatarain-Barrón**, **L. Ricaurte**, **L. Raez**, **M. P. P. Álvarez**, **F. Barrón**, **L. Rojas**, **C. Rolfo**, **N. Karachaliou**, **M. A. Molina-Vila**, and **R. Rosell** (2020). Precision medicine and its implementation in patients with NTRK fusion genes: perspective from developing countries. *Ther. Adv. Respir. Dis.*, **14**, 1753466620938553.

39. **Cescon, D. W.**, **B. Haibe-Kains**, and **T. W. Mak** (2015). APOBEC3B expression in breast cancer reflects cellular proliferation, while a deletion polymorphism is associated with immune activation. *Proc. Natl. Acad. Sci. U. S. A.*, **112**(9), 2841–2846.

40. **Chaiwun, B.** and **P. Thorner** (2007). Fine needle aspiration for evaluation of breast masses. *Curr. Opin. Obstet. Gynecol.*, **19**(1), 48–55.

41. **Chawla, S.**, **S. Samydurai**, **S. L. Kong**, **Z. Wu**, **Z. Wang**, **W. L. Tam**, **D. Sengupta**, and **V. Kumar** (2021). UniPath: a uniform approach for pathway and gene-set based analysis of heterogeneity in single-cell epigenome and transcriptome profiles. *Nucleic Acids Res.*, **49**(3), 1801.

42. **Chen, L.**, **Y. Zeng**, and **S.-F. Zhou** (2018). Role of apoptosis in cancer resistance to chemotherapy. *Current understanding of apoptosis-programmed cell death*.

43. **Chen, L.**, **Y. Zhai**, **Q. He**, **W. Wang**, and **M. Deng** (2020). Integrating deep supervised, Self-Supervised and unsupervised learning for Single-Cell RNA-seq clustering and annotation. *Genes*, **11**(7).

44. **Chen, Y. H.** and **M. C. Magalhaes** (1990). Hypoalbuminemia in patients with multiple myeloma. *Arch. Intern. Med.*, **150**(3), 605–610.

45. **Cheng, J.-Z.**, **D. Ni**, **Y.-H. Chou**, **J. Qin**, **C.-M. Tiu**, **Y.-C. Chang**, **C.-S. Huang**, **D. Shen**, and **C.-M. Chen** (2016). Computer-Aided diagnosis with deep learning architecture: Applications to breast lesions in US images and pulmonary nodules in CT scans. *Sci. Rep.*, **6**, 24454.

46. **Cheng, Y.-H.**, **Y.-C. Chen**, **E. Lin**, **R. Brien**, **S. Jung**, **Y.-T. Chen**, **W. Lee**, **Z. Hao**, **S. Sahoo**, **H. Min Kang**, **J. Cong**, **M. Burness**, **S. Nagrath**, **M. S Wicha**, and **E. Yoon** (2019). Hydro-Seq enables contamination-free high-throughput single-cell RNA-sequencing for circulating tumor cells. *Nat. Commun.*, **10**(1), 2163.

47. **Child, J. A.**, **G. J. Morgan**, **F. E. Davies**, **R. G. Owen**, **S. E. Bell**, **K. Hawkins**, **J. Brown**, **M. T. Drayson**, and **P. J. Selby** (2003). High-dose chemotherapy with hematopoietic stem-cell rescue for multiple myeloma. *N. Engl. J. Med.*, **348**(19), 1875–1883.

48. **Chiu, T.-K.**, **W.-P. Chou**, **S.-B. Huang**, **H.-M. Wang**, **Y.-C. Lin**, **C.-H. Hsieh**, and **M.-H. Wu** (2016). Application of optically-induced-dielectrophoresis in microfluidic system for purification of circulating tumour cells for gene expression analysis- cancer cell line model. *Sci. Rep.*, **6**, 32851.

49. **Christensen, E.**, **A. Naidas**, **D. Chen**, **M. Husic**, and **P. Shooshtari** (2022). TMExplorer: A tumour microenvironment single-cell RNAseq database and search tool. *PLoS One*, **17**(9), e0272302.

50. **Christoph Röllig, M. B., Stefan Knop** (2015). Multiple myeloma. *Lancet*, **385**(9983), 2197–2208.

51. **Chuo, C. B.** and **A. P. Corder** (2003). Core biopsy vs fine needle aspiration cytology in a symptomatic breast clinic. *Eur. J. Surg. Oncol.*, **29**(4), 374–378.

52. **Clark, N. C.**, **A. M. Friel**, **C. A. Pru**, **L. Zhang**, **T. Shioda**, **B. R. Rueda**, **J. J. Peluso**, and **J. K. Pru** (2016). Progesterone receptor membrane component 1 promotes survival of human breast cancer cells and the growth of xenograft tumors. *Cancer Biol. Ther.*, **17**(3), 262–271.

53. **Clark, T. G.**, **M. J. Bradburn**, **S. B. Love**, and **D. G. Altman** (2003). Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, **89**(2), 232–238.

54. **Computing** (). R: A language and environment for statistical computing. *Vienna: R Core Team*.

55. **Coresh, J.**, **T. C. Turin**, **K. Matsushita**, **Y. Sang**, **S. H. Ballew**, **L. J. Appel**, **H. Arima**, **S. J. Chadban**, **M. Cirillo**, **O. Djurdjev**, **J. A. Green**, **G. H. Heine**, **L. A. Inker**, **F. Irie**, **A. Ishani**, **J. H. Ix**, **C. P. Kovesdy**, **A. Marks**, **T. Ohkubo**, **V. Shalev**, **A. Shankar**, **C. P. Wen**, **P. E. de Jong**, **K. Iseki**, **B. Stengel**, **R. T. Gansevoort**, and **A. S. Levey** (2014). Decline in estimated glomerular filtration rate and subsequent risk of end-stage renal disease and mortality. *JAMA*, **311**(24), 2518–2531.

56. **Couturier, C. P.**, **S. Ayyadhury**, **P. U. Le**, **J. Nadaf**, **J. Monlong**, **G. Riva**, **R. Allache**, **S. Baig**, **X. Yan**, **M. Bourgey**, **C. Lee**, **Y. C. D. Wang**, **V. Wee Yong**, **M.-C. Guiot**, **H. Najafabadi**, **B. Misic**, **J. Antel**, **G. Bourque**, **J. Ragoussis**, and **K. Petrecca** (2020). Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. *Nat. Commun.*, **11**(1), 3406.

57. **Cowan, A. J.**, **D. J. Green**, **M. Kwok**, **S. Lee**, **D. G. Coffey**, **L. A. Holmberg**, **S. Tuazon**, **A. K. Gopal**, and **E. N. Libby** (2022). Diagnosis and management of multiple myeloma: A review. *JAMA*, **327**(5), 464–477.

58. **Cristofanilli, M.**, **G. T. Budd**, **M. J. Ellis**, **A. Stopeck**, **J. Matera**, **M. C. Miller**, **J. M. Reuben**, **G. V. Doyle**, **W. J. Allard**, **L. W. M. M. Terstappen**, and **D. F. Hayes** (2004). Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *N. Engl. J. Med.*, **351**(8), 781–791.

59. **Danila, D. C.**, **G. Heller**, **G. A. Gignac**, **R. Gonzalez-Espinoza**, **A. Anand**, **E. Tanaka**, **H. Lilja**, **L. Schwartz**, **S. Larson**, **M. Fleisher**, and **H. I. Scher** (2007). Circulating tumor cell number and prognosis in progressive castration-resistant prostate cancer. *Clin. Cancer Res.*, **13**(23), 7053–7058.

60. **Dare, A. J.**, **B. O. Anderson**, **R. Sullivan**, **C. S. Pramesh**, **I. Andre**, **I. F. Adewole**, **R. A. Badwe**, and **C. L. Gauvreau**, Surgical services for cancer care. The International Bank for Reconstruction and Development / The World Bank, Washington (DC), 2015, 223.

61. **Das, S.**, **A. Rai**, and **S. N. Rai** (2022). Differential expression analysis of Single-Cell RNA-Seq data: Current statistical approaches and outstanding challenges. *Entropy*, **24**(7).

62. **Debela, D. T.**, **S. G. Muzazu**, **K. D. Heraro**, **M. T. Ndalama**, **B. W. Mesele**, **D. C. Haile**, **S. K. Kitui**, and **T. Manyazewal** (2021). New approaches and procedures for cancer treatment: Current perspectives. *SAGE Open Med*, **9**, 20503121211034366.

63. **DeFronzo, R. A.**, **C. R. Cooke**, **J. R. Wright**, and **R. L. Humphrey** (1978). Renal function in patients with multiple myeloma. *Medicine*, **57**(2), 151–166.

64. **Deli, T.**, **M. Orosz**, and **A. Jakab** (2020). Hormone replacement therapy in cancer survivors - review of the literature. *Pathol. Oncol. Res.*, **26**(1), 63–78.

65. **Dellay, B.** and **M. Groth** (2016). Emergency management of Malignancy-Associated hypercalcemia. *Adv. Emerg. Nurs. J.*, **38**(1), 15–25; quiz E1.

66. **Desai, K.**, **J. M. McManus**, and **N. Sharifi** (2021). Hormonal therapy for prostate cancer. *Endocr. Rev.*, **42**(3), 354–373.

67. **Descamps, L.**, **D. Le Roy**, and **A.-L. Deman** (2022). Microfluidic-Based technologies for CTC isolation: A review of 10 years of intense efforts towards liquid biopsy. *Int. J. Mol. Sci.*, **23**(4).

68. **Deyo, R. A.**, **D. C. Cherkin**, **J. Weinstein**, **J. Howe**, **M. Ciol**, and **A. G. Mulley, Jr** (2000). Involving patients in clinical decisions: impact of an interactive video program on use of back surgery. *Med. Care*, **38**(9), 959–969.

69. **Dill, E. A.**, **P. M. Dillon**, **T. N. Bullock**, and **A. M. Mills** (2018). IDO expression in breast cancer: an assessment of 281 primary and metastatic cases with comparison to PD-L1. *Mod. Pathol.*, **31**(10), 1513–1522.

70. **Dimopoulos, M.**, **E. Kastritis**, **L. Rosinol**, **J. Bladé**, and **H. Ludwig** (2008). Pathogenesis and treatment of renal failure in multiple myeloma. *Leukemia*, **22**(8), 1485–1493.

71. **Dimopoulos, M. A.**, **E. Terpos**, **A. Chanan-Khan**, **N. Leung**, **H. Ludwig**, **S. Jagannath**, **R. Niesvizky**, **S. Giralt**, **J.-P. Fermand**, **J. Bladé**, **R. L. Comenzo**, **O. Sezer**, **A. Palumbo**, **J.-L. Harousseau**, **P. G. Richardson**, **B. Barlogie**, **K. C. Anderson**, **P. Sonneveld**, **P. Tosi**, **M. Cavo**, **S. V. Rajkumar**, **B. G. M. Durie**, and **J. San Miguel** (2010). Renal impairment in patients with multiple myeloma: a consensus statement on behalf of the international myeloma working group. *J. Clin. Oncol.*, **28**(33), 4976–4984.

72. **Ding, H.**, **A. Blair**, **Y. Yang**, and **J. M. Stuart** (2019). Biological process activity transformation of single cell gene expression for cross-species alignment. *Nat. Commun.*, **10**(1), 4899.

73. **Ding, J.**, **X. Adiconis**, **S. K. Simmons**, **M. S. Kowalczyk**, **C. C. Hession**, **N. D. Marjanovic**, **T. K. Hughes**, **M. H. Wadsworth**, **T. Burks**, **L. T. Nguyen**, **J. Y. H. Kwon**, **B. Barak**, **W. Ge**, **A. J. Kedaigle**, **S. Carroll**, **S. Li**, **N. Hacohen**, **O. Rozenblatt-Rosen**, **A. K. Shalek**, **A.-C. Villani**, **A. Regev**, and **J. Z. Levin** (2020). Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.*, **38**(6), 737–746.

74. **Dosani, T.**, **F. Covut**, **R. Beck**, **J. J. Driscoll**, **M. de Lima**, and **E. Malek** (2017). Significance of the absolute lymphocyte/monocyte ratio as a prognostic immune biomarker in newly diagnosed multiple myeloma. *Blood Cancer J.*, **7**(6), e579.

75. **Drake, T. M.**, **S. R. Knight**, **E. M. Harrison**, and **K. Søreide** (2018). Global inequities in precision medicine and molecular cancer research. *Front. Oncol.*, **8**, 346.

76. **Durante, M. A.**, **D. A. Rodriguez**, **S. Kurtenbach**, **J. N. Kuznetsov**, **M. I. Sanchez**, **C. L. Decatur**, **H. Snyder**, **L. G. Feun**, **A. S. Livingstone**, and **J. W. Harbour** (2020). Single-cell analysis reveals new evolutionary complexity in uveal melanoma. *Nat. Commun.*, **11**(1), 496.

77. **Durie, B. G. M.** and **S. E. Salmon** (1975). A clinical staging system for multiple myeloma correlation of measured myeloma cell mass with presenting clinical features, response to treatment, and survival. *Cancer*, **36**(3), 842–854.

78. **Ebright, R. Y.**, **S. Lee**, **B. S. Wittner**, **K. L. Niederhoffer**, **B. T. Nicholson**, **A. Bardia**, **S. Truesdell**, **D. F. Wiley**, **B. Wesley**, **S. Li**, **A. Mai**, **N. Aceto**, **N. Vincent-Jordan**, **A. Szabolcs**, **B. Chirn**, **J. Kreuzer**, **V. Comaills**, **M. Kalinich**, **W. Haas**, **D. T. Ting**, **M. Toner**, **S. Vasudevan**, **D. A. Haber**, **S. Maheswaran**, and **D. S. Micalizzi** (2020). Deregulation of ribosomal protein expression and translation promotes breast cancer metastasis. *Science*, **367**(6485), 1468–1473.

79. **Eelen, G.**, **I. Vanden Bempt**, **L. Verlinden**, **M. Drijkoningen**, **A. Smeets**, **P. Neven**, **M. R. Christiaens**, **K. Marchal**, **R. Bouillon**, and **A. Verstuyf** (2008). Expression of the BRCA1-interacting protein Brip1/BACH1/FANCJ is driven by E2F and correlates with human breast cancer malignancy. *Oncogene*, **27**(30), 4233–4241.

80. **Ehrenstein, V.**, **H. Kharrazi**, **H. Lehmann**, and **C. O. Taylor**, Obtaining data from electronic health records. *In Tools and technologies for registry interoperability, registries for evaluating patient outcomes: A user's guide, 3rd edition, Addendum 2 [Internet]*. Agency for Healthcare Research and Quality (US), 2019.

81. **Eloubeidi, M. A.**, **V. K. Chen**, **I. A. Eltoum**, **D. Jhala**, **D. C. Chhieng**, **N. Jhala**, **S. M. Vickers**, and **C. M. Wilcox** (2003). Endoscopic ultrasound–guided fine needle aspiration biopsy of patients with suspected pancreatic cancer: diagnostic accuracy and acute and 30-day complications. *The American journal of gastroenterology*, **98**(12), 2663–2668.

82. **Elston, D. M.**, **E. J. Stratman**, and **S. J. Miller** (2016). Skin biopsy: Biopsy issues in specific diseases. *J. Am. Acad. Dermatol.*, **74**(1), 1–16; quiz 17–8.

83. **Fa, B.**, **T. Wei**, **Y. Zhou**, **L. Johnston**, **X. Yuan**, **Y. Ma**, **Y. Zhang**, and **Z. Yu** (2021). GapClust is a light-weight approach distinguishing rare cells from voluminous single cell expression profiles. *Nat. Commun.*, **12**(1), 4197.

84. **Farace, F.**, **C. Massard**, **N. Vimond**, **F. Drusch**, **N. Jacques**, **F. Billiot**, **A. Laplanche**, **A. Chauchereau**, **L. Lacroix**, **D. Planchard**, **S. Le Moulec**, **F. André**, **K. Fizazi**, **J. C. Soria**, and **P. Vielh** (2011). A direct comparison of CellSearch and ISET for circulating

tumour-cell detection in patients with metastatic carcinomas. *Br. J. Cancer*, **105**(6), 847–853.

85. **Fard, M. M.**, **T. Thonet**, and **E. Gaussier** (2020). Deep k-means: Jointly clustering with k-means and learning representations.

86. **Farris, J. C.**, **A. Ritter**, **M. D. Craig**, **N. Shah**, **L. Veltri**, **A. S. Kanate**, **K. Ross**, and **J. A. Vargo** (2019). Patterns of relapse after salvage autologous stem cell transplant for hodgkin's lymphoma: Should sites of relapse relative to initially involved sites be used to guide indications for peri-transplant radiation therapy. *Practical radiation oncology*, **9**(3), e290–e297.

87. **Feng, S.-S.** and **S. Chien** (2003). Chemotherapeutic engineering: Application and further development of chemical engineering principles for chemotherapy of cancer and other diseases. *Chem. Eng. Sci.*, **58**(18), 4087–4114.

88. **Ferlay, J.**, **M. Colombet**, **I. Soerjomataram**, **D. M. Parkin**, **M. Piñeros**, **A. Znaor**, and **F. Bray** (2021). Cancer statistics for the year 2020: An overview. *Int. J. Cancer*.

89. **Fermand, J.-P.**, **S. Katsahian**, **M. Divine**, **V. Leblond**, **F. Dreyfus**, **M. Macro**, **B. Arnulf**, **B. Royer**, **X. Mariette**, **E. Pertuiset**, and **Others** (2005). High-dose therapy and autologous blood stem-cell transplantation compared with conventional treatment in myeloma patients aged 55 to 65 years: long-term results of a randomized control trial from the group Myelome-Autogreffe. *J. Clin. Oncol.*, **23**(36), 9227–9233.

90. **Ferreira, M. M.**, **V. C. Ramani**, and **S. S. Jeffrey** (2016). Circulating tumor cell technologies. *Mol. Oncol.*, **10**(3), 374–394.

91. **Fitzmaurice, C.**, **C. Allen**, **R. M. Barber**, **L. Barregard**, **Z. A. Bhutta**, **H. Brenner**, **D. J. Dicker**, **O. Chimed-Orchir**, **R. Dandona**, **L. Dandona**, and **Others** (2017). Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. *JAMA oncology*, **3**(4), 524–548.

92. **Follain, G.**, **N. Osmani**, **A. S. Azevedo**, **G. Allio**, **L. Mercier**, **M. A. Karreman**, **G. Solecki**, **M. J. Garcia Leòn**, **O. Lefebvre**, **N. Fekonja**, **C. Hille**, **V. Chabannes**,

G. Dollé, **T. Metivet**, **F. D. Hovsepian**, **C. Prudhomme**, **A. Pichot**, **N. Paul**, **R. Carapito**, **S. Bahram**, **B. Ruthensteiner**, **A. Kemmling**, **S. Siemonsen**, **T. Schneider**, **J. Fiehler**, **M. Glatzel**, **F. Winkler**, **Y. Schwab**, **K. Pantel**, **S. Harlepp**, and **J. G. Goetz** (2018). Hemodynamic forces tune the arrest, adhesion, and extravasation of circulating tumor cells. *Dev. Cell*, **45**(1), 33–52.e12.

93. **Fortunato, O.**, **P. Gasparini**, **M. Boeri**, and **G. Sozzi** (2019). Exo-miRNAs as a new tool for liquid biopsy in lung cancer. *Cancers*, **11**(6).

94. **Friedman, N.**, **D. Geiger**, and **M. Goldszmidt** (1997). Bayesian network classifiers. *Mach. Learn.*, **29**(2), 131–163.

95. **Fu, T.**, **T. N. Hoang**, **C. Xiao**, and **J. Sun**, DDL: Deep dictionary learning for predictive phenotyping. *In IJCAI*. 2019.

96. **Gabriel, M. T.**, **L. R. Calleja**, **A. Chalopin**, **B. Ory**, and **D. Heymann** (2016). Circulating tumor cells: A review of Non–EpCAM-Based approaches for cell enrichment and isolation. *Clin. Chem.*, **62**(4), 571–581.

97. **Gaki, V.**, **M. Tsopanomichalou**, **G. Sourvinos**, **D. Tsiftsis**, and **D. A. Spandidos** (2000). Allelic loss in chromosomal region 1q21–23 in breast cancer is associated with peritumoral angiolymphatic invasion and extensive intraductal component. *Eur. J. Surg. Oncol.*, **26**(5), 455–460.

98. **Gambardella, V.**, **N. Tarazona**, **J. M. Cejalvo**, **P. Lombardi**, **M. Huerta**, **S. Roselló**, **T. Fleitas**, **D. Roda**, and **A. Cervantes** (2020). Personalized medicine: Recent progress in cancer therapy. *Cancers*, **12**(4).

99. **Garbe, C.**, **P. Büttner**, **J. Bertz**, **G. Burg**, **B. d'Hoedt**, **H. Drepper**, **I. Guggenmoos-Holzmann**, **W. Lechner**, **A. Lippold**, and **C. E. Orfanos** (1995). Primary cutaneous melanoma. identification of prognostic groups and estimation of individual prognosis for 5093 patients. *Cancer*, **75**(10), 2484–2491.

100. **Garnis, C.**, **J. Campbell**, **J. J. Davies**, **C. Macaulay**, **S. Lam**, and **W. L. Lam** (2005). Involvement of multiple developmental genes on chromosome 1p in lung tumorigenesis. *Hum. Mol. Genet.*, **14**(4), 475–482.

101. **Gastwirth, J. L.** (1972). The estimation of the lorenz curve and gini index. *Rev. Econ. Stat.*, **54**(3), 306–316.

102. **Gay, F.**, **S. Oliva**, **M. T. Petrucci**, **C. Conticello**, **L. Catalano**, **P. Corradini**, **A. Siniscalchi**, **V. Magarotto**, **L. Pour**, **A. Carella**, and **Others** (2015). Chemotherapy plus lenalidomide versus autologous transplantation, followed by lenalidomide plus prednisone versus lenalidomide maintenance, in patients with multiple myeloma: a randomised, multicentre, phase 3 trial. *Lancet Oncol.*, **16**(16), 1617–1629.

103. **Gay, F.**, **S. Oliva**, **M. T. Petrucci**, **V. Montefusco**, **C. Conticello**, **P. Musto**, **L. Catalano**, **A. Evangelista**, **S. Spada**, **P. Campbell**, and **Others** (2017). Autologous transplant vs oral chemotherapy and lenalidomide in newly diagnosed young myeloma patients: a pooled analysis. *Leukemia*, **31**(8), 1727.

104. **Genc, O.** and **A. Dag** (2016). A bayesian network-based data analytical approach to predict velocity distribution in small streams. *J. Hydroinformatics*, **18**(3), 466–480.

105. **Giannopoulou, L.**, **S. Kasimir-Bauer**, and **E. S. Lianidou** (2018). Liquid biopsy in ovarian cancer: recent advances on circulating tumor cells and circulating tumor DNA. *Clin. Chem. Lab. Med.*, **56**(2), 186–197.

106. **Gilad, Y.**, **G. Gellerman**, **D. M. Lonard**, and **B. W. O'Malley** (2021). Drug combination in cancer Treatment-From cocktails to conjugated combinations. *Cancers*, **13**(4).

107. **Gill, I. S.**, **M. M. Desai**, **J. H. Kaouk**, **A. M. Meraney**, **D. P. Murphy**, **G. T. Sung**, and **A. C. Novick** (2002). Laparoscopic partial nephrectomy for renal tumor: duplicating open surgical techniques. *J. Urol.*, **167**(2 Pt 1), 469–7; discussion 475–6.

108. **Giuliano, M.**, **A. Giordano**, **S. Jackson**, **K. R. Hess**, **U. De Giorgi**, **M. Mego**, **B. C. Handy**, **N. T. Ueno**, **R. H. Alvarez**, **M. De Laurentiis**, **S. De Placido**, **V. Valero**, **G. N. Hortobagyi**, **J. M. Reuben**, and **M. Cristofanilli** (2011). Circulating tumor cells as prognostic and predictive markers in metastatic breast cancer patients receiving first-line systemic treatment. *Breast Cancer Res.*, **13**(3), R67.

109. **Glinsky, G. V.**, **O. Berezovska**, and **A. B. Glinskii** (2005). Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *J. Clin. Invest.*, **115**(6), 1503–1521.

110. **Goel, M. K.**, **P. Khanna**, and **J. Kishore** (2010). Understanding survival analysis: Kaplan-Meier estimate. *Int. J. Ayurveda Res.*, **1**(4), 274–278.

111. **Golden, D. P.** and **J. R. Hooley** (1994). Oral mucosal biopsy procedures. excisional and incisional. *Dent. Clin. North Am.*, **38**(2), 279–300.

112. **Goldstein, L. D.**, **Y.-J. J. Chen**, **J. Dunne**, **A. Mir**, **H. Hubschle**, **J. Guillory**, **W. Yuan**, **J. Zhang**, **J. Stinson**, **B. Jaiswal**, **K. B. Pahuja**, **I. Mann**, **T. Schaal**, **L. Chan**, **S. Anandakrishnan**, **C.-W. Lin**, **P. Espinoza**, **S. Husain**, **H. Shapiro**, **K. Swaminathan**, **S. Wei**, **M. Srinivasan**, **S. Seshagiri**, and **Z. Modrusan** (2017). Massively parallel nanowell-based single-cell gene expression profiling. *BMC Genomics*, **18**(1), 519.

113. **Gole, J.**, **A. Gore**, **A. Richards**, **Y.-J. Chiu**, **H.-L. Fung**, **D. Bushman**, **H.-I. Chiang**, **J. Chun**, **Y.-H. Lo**, and **K. Zhang** (2013). Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nat. Biotechnol.*, **31**(12), 1126–1132.

114. **Gonsalves, W. I.**, **N. Leung**, **S. V. Rajkumar**, **A. Dispenzieri**, **M. Q. Lacy**, **S. R. Hayman**, **F. K. Buadi**, **D. Dingli**, **P. Kapoor**, **R. S. Go**, **Y. Lin**, **S. J. Russell**, **J. A. Lust**, **S. Zeldenrust**, **R. A. Kyle**, **M. A. Gertz**, and **S. K. Kumar** (2015). Improvement in renal function and its impact on survival in patients with newly diagnosed multiple myeloma. *Blood Cancer J.*, **5**(3), e296.

115. **Goswami, C.**, **S. Poonia**, **L. Kumar**, and **D. Sengupta** (2019). Staging system to predict the risk of relapse in multiple myeloma patients undergoing autologous stem cell transplantation. *Front. Oncol.*, **9**, 633.

116. **Greipp, P. R.**, **J. S. Miguel**, **B. G. M. Durie**, **J. J. Crowley**, **B. Barlogie**, **J. Bladé**, **M. Boccadoro**, **J. A. Child**, **H. Avet-Loiseau**, **R. A. Kyle**, and **Others** (2005). International staging system for multiple myeloma. *J. Clin. Oncol.*, **23**(15), 3412–3420.

117. **Guo, M.**, **H. Wang**, **S. S. Potter**, **J. A. Whitsett**, and **Y. Xu** (2015). SINCERA: A pipeline for Single-Cell RNA-Seq profiling analysis. *PLoS Comput. Biol.*, **11**(11), e1004575.

118. **Guo, Y.**, **Y. Dai**, **H. Yu**, **S. Zhao**, **D. C. Samuels**, and **Y. Shyr** (2017). Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, **109**(2), 83–90.

119. **Gupta, A.**, **L. Kumar**, **D. Dabkara**, **D. Gupta**, **O. Sharma**, and **V. Sreeniwas** (2009). Multiple myeloma: Autologous stem cell transplantation versus conventional chemotherapy—a retrospective age and stage matched analysis. *J. Clin. Oncol.*, **27**(15_suppl), 7041–7041.

120. **Gupta, K.**, **S. K. Mohanty**, **A. Mittal**, **S. Kalra**, **S. Kumar**, **T. Mishra**, **J. Ahuja**, **D. Sengupta**, and **G. Ahuja** (2020). The cellular basis of the loss of smell in 2019-nCoV-infected individuals. *Brief. Bioinform.*.

121. **Gutschner, T.** and **S. Diederichs** (2012). The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biol.*, **9**(6), 703–719.

122. **Haber, D. A.** and **V. E. Velculescu** (2014). Blood-based analyses of cancer: circulating tumor cells and circulating tumor DNA. *Cancer Discov.*, **4**(6), 650–661.

123. **Habli, Z.**, **W. AlChamaa**, **R. Saab**, **H. Kadara**, and **M. L. Khraiche** (2020). Circulating tumor cell detection technologies and clinical utility: Challenges and opportunities. *Cancers*, **12**(7).

124. **Haghverdi, L.**, **A. T. L. Lun**, **M. D. Morgan**, and **J. C. Marioni** (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.*, **36**(5), 421–427.

125. **Han, L.**, **J. S. Lee**, **T. Mori**, **H. Zhang**, **G. Landberg**, **A. Kallioniemi**, **P. Argani**, and **S. Sukumar** (2008). HEYL, an overexpressed gene in breast cancer, functions as a novel negative regulator of TGF-β pathway. *Cancer Res.*, **68**(9_Supplement), 5201–5201.

126. **Hanahan, D.** and **R. A. Weinberg** (2000). The hallmarks of cancer. *Cell*, **100**(1), 57–70.

127. **Hanahan, D.** and **R. A. Weinberg** (2011). Hallmarks of cancer: the next generation. *Cell*, **144**(5), 646–674.

128. **Hänzelmann, S.**, **R. Castelo**, and **J. Guinney** (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, **14**, 7.

129. **Hao, Y.**, **S. Hao**, **E. Andersen-Nissen**, **W. M. Mauck, 3rd**, **S. Zheng**, **A. Butler**, **M. J. Lee**, **A. J. Wilk**, **C. Darby**, **M. Zager**, **P. Hoffman**, **M. Stoeckius**, **E. Papalexi**,

E. P. Mimitou, **J. Jain**, **A. Srivastava**, **T. Stuart**, **L. M. Fleming**, **B. Yeung**, **A. J. Rogers**, **J. M. McElrath**, **C. A. Blish**, **R. Gottardo**, **P. Smibert**, and **R. Satija** (2021). Integrated analysis of multimodal single-cell data. *Cell*, **184**(13), 3573–3587.e29.

130. **Haque, A.**, **J. Engel**, **S. A. Teichmann**, and **T. Lönnberg** (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.*, **9**(1), 75.

131. **He, J.**, **W. Tan**, and **J. Ma** (2017). Circulating tumor cells and dna for real-time egfr detection and monitoring of non-small-cell lung cancer. *Future Oncology*, **13**(9), 787–797.

132. **He, Q.**, **X. Fan**, **T. Yuan**, **L. Kong**, **X. Du**, **D. Zhuang**, and **Z. Fan** (2007). Eleven years of experience reveals that fine-needle aspiration cytology is still a useful method for preoperative diagnosis of breast carcinoma. *Breast*, **16**(3), 303–306.

133. **Heckerman, D.**, A tutorial on learning with bayesian networks. *In* **M. I. Jordan** (ed.), *Learning in Graphical Models*. Springer Netherlands, Dordrecht, 1998, 301–354.

134. **Heywang-Köbrunner, S. H.**, **A. Heinig**, **U. Schaumlöffel**, **P. Viehweg**, **J. Buchmann**, **D. Lampe**, and **R.-P. Spielmann** (1999). MR-guided percutaneous excisional and incisional biopsy of breast lesions.

135. **Hirsch, F. R.**, **A. McElhinny**, **D. Stanforth**, **J. Ranger-Moore**, **M. Jansson**, **K. Kulangara**, **W. Richardson**, **P. Towne**, **D. Hanks**, **B. Vennapusa**, **A. Mistry**, **R. Kalamegham**, **S. Averbuch**, **J. Novotny**, **E. Rubin**, **K. Emancipator**, **I. McCaffery**, **J. A. Williams**, **J. Walker**, **J. Longshore**, **M. S. Tsao**, and **K. M. Kerr** (2017). PD-L1 immunohistochemistry assays for lung cancer: Results from phase 1 of the blueprint PD-L1 IHC assay comparison project. *J. Thorac. Oncol.*, **12**(2), 208–222.

136. **Hodgkinson, C. L.**, **C. J. Morrow**, **Y. Li**, **R. L. Metcalf**, **D. G. Rothwell**, **F. Trapani**, **R. Polanski**, **D. J. Burt**, **K. L. Simpson**, **K. Morris**, **S. D. Pepper**, **D. Nonaka**, **A. Greystoke**, **P. Kelly**, **B. Bola**, **M. G. Krebs**, **J. Antonello**, **M. Ayub**, **S. Faulkner**, **L. Priest**, **L. Carter**, **C. Tate**, **C. J. Miller**, **F. Blackhall**, **G. Brady**, and **C. Dive** (2014). Tumorigenicity and genetic profiling of circulating tumor cells in small-cell lung cancer. *Nat. Med.*, **20**(8), 897–903.

137. **Hong, Y.**, **F. Fang**, and **Q. Zhang** (2016). Circulating tumor cell clusters: What we know and what we expect (review).

138. **Howe, K. L.**, **P. Achuthan**, **J. Allen**, **J. Allen**, **J. Alvarez-Jarreta**, **M. R. Amode**, **I. M. Armean**, **A. G. Azov**, **R. Bennett**, **J. Bhai**, **K. Billis**, **S. Boddu**, **M. Charkhchi**, **C. Cummins**, **L. Da Rin Fioretto**, **C. Davidson**, **K. Dodiya**, **B. El Houdaigui**, **R. Fatima**, **A. Gall**, **C. Garcia Giron**, **T. Grego**, **C. Guijarro-Clarke**, **L. Haggerty**, **A. Hemrom**, **T. Hourlier**, **O. G. Izuogu**, **T. Juettemann**, **V. Kaikala**, **M. Kay**, **I. Lavidas**, **T. Le**, **D. Lemos**, **J. Gonzalez Martinez**, **J. C. Marugán**, **T. Maurel**, **A. C. McMahon**, **S. Mohanan**, **B. Moore**, **M. Muffato**, **D. N. Oheh**, **D. Paraschas**, **A. Parker**, **A. Parton**, **I. Prosovetskaia**, **M. P. Sakthivel**, **A. I. A. Salam**, **B. M. Schmitt**, **H. Schuilenburg**, **D. Sheppard**, **E. Steed**, **M. Szpak**, **M. Szuba**, **K. Taylor**, **A. Thormann**, **G. Threadgold**, **B. Walts**, **A. Winterbottom**, **M. Chakiachvili**, **A. Chaubal**, **N. De Silva**, **B. Flint**, **A. Frankish**, **S. E. Hunt**, **G. R. IIsley**, **N. Langridge**, **J. E. Loveland**, **F. J. Martin**, **J. M. Mudge**, **J. Morales**, **E. Perry**, **M. Ruffier**, **J. Tate**, **D. Thybert**, **S. J. Trevanion**, **F. Cunningham**, **A. D. Yates**, **D. R. Zerbino**, and **P. Flicek** (2021). Ensembl 2021. *Nucleic Acids Res.*, **49**(D1), D884–D891.

139. **Hu, Z.**, **S. A. Brooks**, **V. Dormoy**, **C.-W. Hsu**, **H.-Y. Hsu**, **L.-T. Lin**, **T. Massfelder**, **W. K. Rathmell**, **M. Xia**, **F. Al-Mulla**, **R. Al-Temaimi**, **A. Amedei**, **D. G. Brown**, **K. R. Prudhomme**, **A. Colacci**, **R. A. Hamid**, **C. Mondello**, **J. Raju**, **E. P. Ryan**, **J. Woodrick**, **A. I. Scovassi**, **N. Singh**, **M. Vaccari**, **R. Roy**, **S. Forte**, **L. Memeo**, **H. K. Salem**, **L. Lowe**, **L. Jensen**, **W. H. Bisson**, and **N. Kleinstreuer** (2015). Assessing the carcinogenic potential of low-dose exposures to chemical mixtures in the environment: focus on the cancer hallmark of tumor angiogenesis. *Carcinogenesis*, **36 Suppl 1**(Suppl 1), S184–202.

140. **Huang, B.**, **J. Li**, **J. Lu**, **Y. Xiao**, **Y. Zhao**, and **H. Huang** (2017). Scoring system for predicting risk of relapse in patients with multiple myeloma within two years after stem cell transplantation. *Blood*, **130**(Suppl 1), 4531–4531.

141. **Huang, S.**, **J. Yang**, **S. Fong**, and **Q. Zhao** (2020). Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Lett.*, **471**, 61–71.

142. **Hutchison, C. A.**, **V. Batuman**, **J. Behrens**, **F. Bridoux**, **C. Sirac**, **A. Dispenzieri**, **G. A. Herrera**, **H. Lachmann**, **P. W. Sanders**, and **International Kidney and Mon-**

oclonal Gammopathy Research Group (2011). The pathogenesis and diagnosis of acute kidney injury in multiple myeloma. *Nat. Rev. Nephrol.*, **8**(1), 43–51.

143. **Ignatiadis, M.** and **M. Reinholz** (2011). Minimal residual disease and circulating tumor cells in breast cancer. *Breast Cancer Res.*, **13**(5), 222.

144. **Ignatiadis, M.**, **G. W. Sledge**, and **S. S. Jeffrey** (2021). Liquid biopsy enters the clinic—implementation issues and future challenges. *Nat. Rev. Clin. Oncol.*, **18**(5), 297–312.

145. **Ilie, M.**, **V. Hofman**, **E. Long-Mira**, **E. Selva**, **J.-M. Vignaud**, **B. Padovani**, **J. Mouroux**, **C.-H. Marquette**, and **P. Hofman** (2014). "sentinel" circulating tumor cells allow early diagnosis of lung cancer in patients with chronic obstructive pulmonary disease. *PLoS One*, **9**(10), e111597.

146. **Institute, N. C.** and **National Cancer Institute** (2020). Cancer diagnosis.

147. **Issa, Z. A.**, **M. S. Zantout**, and **S. T. Azar** (2011). Multiple myeloma and diabetes. *ISRN Endocrinol.*, **2011**, 815013.

148. **Iyer, A.**, **K. Gupta**, **S. Sharma**, **K. Hari**, **Y. F. Lee**, **N. Ramalingam**, **Y. S. Yap**, **J. West**, **A. A. Bhagat**, **B. V. Subramani**, **B. Sabuwala**, **T. Z. Tan**, **J. P. Thiery**, **M. K. Jolly**, **N. Ramalingam**, and **D. Sengupta** (2020). Integrative analysis and machine learning based characterization of single circulating tumor cells. *J. Clin. Med. Res.*, **9**(4).

149. **Jasti, V. D. P.**, **A. S. Zamani**, **K. Arumugam**, **M. Naved**, **H. Pallathadka**, **F. Sammy**, **A. Raghuvanshi**, and **K. Kaliyaperumal** (2022). Computational technique based on machine learning and image processing for medical image analysis of breast cancer diagnosis. *Security and communication networks*, **2022**, 1–7.

150. **Jia, S.**, **R. Zhang**, **Z. Li**, and **J. Li** (2017). Clinical and biological significance of circulating tumor cells, circulating tumor DNA, and exosomes as biomarkers in colorectal cancer. *Oncotarget*, **8**(33), 55632–55645.

151. **Jiang, Y.**, **Y. Qiu**, **A. J. Minn**, and **N. R. Zhang** (2016). Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, **113**(37), E5528–37.

152. **Jin, K.-T.**, **X.-Y. Chen**, **H.-R. Lan**, **S.-B. Wang**, **X.-J. Ying**, **S. M. Abdi**, **W. Wang**, **Z.-M. Hu**, and **X.-Z. Mou** (2019*a*). Current progress in the clinical use of circulating tumor cells as prognostic biomarkers. *Cancer Cytopathol.*, **127**(12), 739–749.

153. **Jin, L.**, **X.-Y. Zuo**, **W.-Y. Su**, **X.-L. Zhao**, **M.-Q. Yuan**, **L.-Z. Han**, **X. Zhao**, **Y.-D. Chen**, and **S.-Q. Rao** (2014). Pathway-based analysis tools for complex diseases: a review. *Genomics Proteomics Bioinformatics*, **12**(5), 210–220.

154. **Jin, X.**, **Y. Li**, **Y. Guo**, **Y. Jia**, **H. Qu**, **Y. Lu**, **P. Song**, **X. Zhang**, **Y. Shao**, **D. Qi**, **W. Xu**, and **C. Quan** (2019*b*). ERα is required for suppressing OCT4-induced proliferation of breast cancer cells via DNMT1/ISL1/ERK axis. *Cell Prolif.*, **52**(4), e12612.

155. **Jordan, N. V.**, **A. Bardia**, **B. S. Wittner**, **C. Benes**, **M. Ligorio**, **Y. Zheng**, **M. Yu**, **T. K. Sundaresan**, **J. A. Licausi**, **R. Desai**, **R. M. O'Keefe**, **R. Y. Ebright**, **M. Boukhali**, **S. Sil**, **M. L. Onozato**, **A. J. Iafrate**, **R. Kapur**, **D. Sgroi**, **D. T. Ting**, **M. Toner**, **S. Ramaswamy**, **W. Haas**, **S. Maheswaran**, and **D. A. Haber** (2016). HER2 expression identifies dynamic functional states within circulating breast cancer cells. *Nature*, **537**(7618), 102–106.

156. **Jovic, D.**, **X. Liang**, **H. Zeng**, **L. Lin**, **F. Xu**, and **Y. Luo** (2022). Single-cell RNA sequencing technologies and applications: A brief overview. *Clin. Transl. Med.*, **12**(3), e694.

157. **Ju, S.**, **C. Chen**, **J. Zhang**, **L. Xu**, **X. Zhang**, **Z. Li**, **Y. Chen**, **J. Zhou**, **F. Ji**, and **L. Wang** (2022). Detection of circulating tumor cells: opportunities and challenges. *Biomarker research*, **10**(1), 1–25.

158. **Kamal, M.**, **W. Razaq**, **M. Leslie**, **S. Adhikari**, and **T. Tanaka**, Circulating tumor cells in breast cancer: A potential liquid biopsy. *In* **P. Van Pham** (ed.), *Breast Cancer*, chapter 6. IntechOpen, Rijeka, 2017.

159. **Kanwar, N.**, **P. Hu**, **P. Bedard**, **M. Clemons**, **D. McCready**, and **S. J. Done** (2015). Identification of genomic signatures in circulating tumor cells from breast cancer. *Int. J. Cancer*, **137**(2), 332–344.

160. **Kaplan, E. L.** and **P. Meier** (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, **53**(282), 457–481.

161. **Karolchik, D.**, **R. Baertsch**, **M. Diekhans**, **T. S. Furey**, **A. Hinrichs**, **Y. T. Lu**, **K. M. Roskin**, **M. Schwartz**, **C. W. Sugnet**, **D. J. Thomas**, **R. J. Weber**, **D. Haussler**, **W. J. Kent**, and **University of California Santa Cruz** (2003). The UCSC genome browser database. *Nucleic Acids Res.*, **31**(1), 51–54.

162. **Katagiri, D.**, **E. Noiri**, and **F. Hinoshita** (2013). Multiple myeloma and kidney disease. *ScientificWorldJournal*, **2013**, 487285.

163. **Kazandjian, D.** (2016). Multiple myeloma epidemiology and survival: A unique malignancy. *Semin. Oncol.*, **43**(6), 676–681.

164. **Kelly, S.**, **F. Humby**, **A. Filer**, **N. Ng**, **M. Di Cicco**, and **others** (2015). Ultrasound-guided synovial biopsy: a safe, well-tolerated and reliable technique for obtaining high-quality synovial tissue from both large and small joints in early .... *Annals of the*.

165. **Kim, C.**, **R. Gao**, **E. Sei**, **R. Brandt**, **J. Hartman**, **T. Hatschek**, **N. Crosetto**, **T. Foukakis**, and **N. E. Navin** (2018). Chemoresistance evolution in Triple-Negative breast cancer delineated by Single-Cell sequencing. *Cell*, **173**(4), 879–893.e13.

166. **Kim, D.**, **G. Pertea**, **C. Trapnell**, **H. Pimentel**, **R. Kelley**, and **S. L. Salzberg** (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**(4), R36.

167. **Kim, J. E.**, **C. Yoo**, **D. H. Lee**, **S.-W. Kim**, **J.-S. Lee**, and **C. Suh** (2010). Serum albumin level is a significant prognostic factor reflecting disease severity in symptomatic multiple myeloma. *Annals of hematology*, **89**, 391–397.

168. **Kiselev, V. Y.**, **T. S. Andrews**, and **M. Hemberg** (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.*, **20**(5), 273–282.

169. **Kiselev, V. Y.**, **K. Kirschner**, **M. T. Schaub**, **T. Andrews**, **A. Yiu**, **T. Chandra**, **K. N. Natarajan**, **W. Reik**, **M. Barahona**, **A. R. Green**, and **M. Hemberg** (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods*, **14**(5), 483–486.

170. **Kitazawa, R.**, **S. Kitazawa**, **K. Kajimoto**, **H. Sowa**, **T. Sugimoto**, **T. Matsui**, **K. Chihara**, and **S. Maeda** (2002). Expression of parathyroid hormone-related protein (PTHrP) in multiple myeloma. *Pathol. Int.*, **52**(1), 63–68.

171. **Knudsen, L. M.**, **M. Hjorth**, **E. Hippe**, and **Nordic Myeloma Study Group** (2000). Renal failure in multiple myeloma: reversibility and impact on the prognosis. *Eur. J. Haematol.*, **65**(3), 175–181.

172. **Koch, C.**, **A. Kuske**, **S. A. Joosse**, **G. Yigit**, **G. Sflomos**, **S. Thaler**, **D. J. Smit**, **S. Werner**, **K. Borgmann**, **S. Gärtner**, **P. Mossahebi Mohammadi**, **L. Battista**, **L. Cayrefourcq**, **J. Altmüller**, **G. Salinas-Riester**, **K. Raithatha**, **A. Zibat**, **Y. Goy**, **L. Ott**, **K. Bartkowiak**, **T. Z. Tan**, **Q. Zhou**, **M. R. Speicher**, **V. Müller**, **T. M. Gorges**, **M. Jücker**, **J.-P. Thiery**, **C. Brisken**, **S. Riethdorf**, **C. Alix-Panabières**, and **K. Pantel** (2020). Characterization of circulating breast cancer cells with tumorigenic and metastatic capacity. *EMBO Mol. Med.*, **12**(9), e11908.

173. **Kooi, T.**, **G. Litjens**, **B. van Ginneken**, **A. Gubern-Mérida**, **C. I. Sánchez**, **R. Mann**, **A. den Heeten**, and **N. Karssemeijer** (2017). Large scale deep learning for computer aided detection of mammographic lesions. *Med. Image Anal.*, **35**, 303–312.

174. **Korsunsky, I.**, **N. Millard**, **J. Fan**, **K. Slowikowski**, **F. Zhang**, **K. Wei**, **Y. Baglaenko**, **M. Brenner**, **P.-R. Loh**, and **S. Raychaudhuri** (2019). Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods*, **16**(12), 1289–1296.

175. **Kowalik, A.**, **M. Kowalewska**, and **S. Góźdź** (2017). Current approaches for avoiding the limitations of circulating tumor cells detection methods-implications for diagnosis and treatment of patients with solid tumors. *Transl. Res.*, **185**, 58–84.e15.

176. **Kraisangka, J.** and **M. J. Druzdzel** (2018). A bayesian network interpretation of the cox's proportional hazard model. *Int. J. Approx. Reason.*, **103**, 195–211.

177. **Krämer, A.**, **J. Green**, **J. Pollard, Jr**, and **S. Tugendreich** (2014). Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, **30**(4), 523–530.

178. **Krebs, M. G.**, **R. L. Metcalf**, **L. Carter**, **G. Brady**, **F. H. Blackhall**, and **C. Dive** (2014). Molecular analysis of circulating tumour cells—biology and biomarkers. *Nat. Rev. Clin. Oncol.*, **11**(3), 129–144.

179. **Kropff, M. H.**, **N. Lang**, **G. Bisping**, **N. Dominé**, **G. Innig**, **M. Hentrich**, **M. Mitterer**, **T. Südhoff**, **R. Fenk**, **C. Straka**, **A. Heinecke**, **O. M. Koch**, **H. Ostermann**, **W. E. Berdel**, and **J. Kienast** (2003). Hyperfractionated cyclophosphamide in combi-

nation with pulsed dexamethasone and thalidomide (HyperCDT) in primary refractory or relapsed multiple myeloma. *Br. J. Haematol.*, **122**(4), 607–616.

180. **Kuhn, M.** (2015). caret: Classification and regression training.

181. **Kumar, L.**, **R. R. Boya**, **R. Pai**, **P. Harish**, **A. Mookerjee**, **B. Sainath**, **M. B. Patekar**, **R. K. Sahoo**, **P. S. Malik**, **O. Sharma**, *et al.* (2016). Autologous stem cell transplantation for multiple myeloma: Long-term results. *National Medical Journal of India*, **29**(4).

182. **Kumar, L.**, **J. Ghosh**, **P. Ganessan**, **A. Gupta**, **R. Hariprasad**, and **V. Kochupillai** (2009). High-dose chemotherapy with autologous stem cell transplantation for multiple myeloma: what predicts the outcome? experience from a developing country. *Bone Marrow Transplant.*, **43**(6), 481.

183. **Kumar, S.**, **L. Baizer**, **N. S. Callander**, **S. A. Giralt**, **J. Hillengass**, **B. Freidlin**, **A. Hoering**, **P. G. Richardson**, **E. I. Schwartz**, **A. Reiman**, **S. Lentzsch**, **P. L. McCarthy**, **S. Jagannath**, **A. J. Yee**, **R. F. Little**, and **N. S. Raje** (2022). Gaps and opportunities in the treatment of relapsed-refractory multiple myeloma: Consensus recommendations of the NCI multiple myeloma steering committee. *Blood Cancer J.*, **12**(6), 98.

184. **Kumar, S. K.**, **S. V. Rajkumar**, **A. Dispenzieri**, **M. Q. Lacy**, **S. R. Hayman**, **F. K. Buadi**, **S. R. Zeldenrust**, **D. Dingli**, **S. J. Russell**, **J. A. Lust**, and **Others** (2008). Improved survival in multiple myeloma and the impact of novel therapies. *Blood*, **111**(5), 2516–2520.

185. **Kursa, M. B.** and **W. R. Rudnicki** (2010). Feature selection with the boruta package. *J. Stat. Softw.*, **36**, 1–13.

186. **Kwa, M.** and **F. J. Esteva** (2018). Detection and clinical implications of occult systemic micrometastatic breast cancer.

187. **Kyle, R. A.** and **S. V. Rajkumar** (2004). Multiple myeloma. *N. Engl. J. Med.*, **351**(18), 1860–1873.

188. **Lacny, S.**, **T. Wilson**, **F. Clement**, **D. J. Roberts**, **P. D. Faris**, **W. A. Ghali**, and **D. A. Marshall** (2015). Kaplan-Meier survival analysis overestimates the risk of revision arthroplasty: A meta-analysis. *Clin. Orthop. Relat. Res.*, **473**(11), 3431–3442.

189. **Ladyzynski, P.**, **M. Molik**, and **P. Foltynski** (2022). Dynamic bayesian networks for prediction of health status and treatment effect in patients with chronic lymphocytic leukemia. *Sci. Rep.*, **12**(1), 1811.

190. **Lan, L.**, **Y. Luo**, **M. Zhou**, **L. Huo**, **H. Chen**, **Q. Zuo**, and **W. Deng** (2020). Comparison of diagnostic accuracy of thyroid cancer with Ultrasound-Guided Fine-Needle aspiration and Core-Needle biopsy: A systematic review and Meta-Analysis. *Front. Endocrinol.*, **11**, 44.

191. **Langmead, B.** (2010). Aligning short sequencing reads with bowtie. *Curr. Protoc. Bioinformatics*, **Chapter 11**, Unit 11.7.

192. **Langmead, B.** and **S. L. Salzberg** (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods*, **9**(4), 357–359.

193. **Larkin, J.**, **V. Chiarion-Sileni**, **R. Gonzalez**, **J. J. Grob**, **C. L. Cowey**, **C. D. Lao**, **D. Schadendorf**, **R. Dummer**, **M. Smylie**, **P. Rutkowski**, *et al.* (2015). Combined nivolumab and ipilimumab or monotherapy in untreated melanoma. *New England journal of medicine*, **373**(1), 23–34.

194. **Law, C. W.**, **Y. Chen**, **W. Shi**, and **G. K. Smyth** (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**(2), R29.

195. **Lee, Y.**, **G. Guan**, and **A. A. Bhagat** (2018). ClearCell® FX, a label-free microfluidics technology for enrichment of viable circulating tumor cells. *Cytometry A*, **93**(12), 1251–1254.

196. **Lehners, N.**, **N. Becker**, **A. Benner**, **M. Pritsch**, **M. Löpprich**, **E. K. Mai**, **J. Hillengass**, **H. Goldschmidt**, and **M.-S. Raab** (2018). Analysis of long-term survival in multiple myeloma after first-line autologous stem cell transplantation: impact of clinical risk factors and sustained response. *Cancer Med.*, **7**(2), 307–316.

197. **Lei, Y.**, **R. Tang**, **J. Xu**, **W. Wang**, **B. Zhang**, **J. Liu**, **X. Yu**, and **S. Shi** (2021). Applications of single-cell sequencing in cancer research: progress and perspectives. *J. Hematol. Oncol.*, **14**(1), 91.

198. **Levine, D. S.**, **R. C. Haggitt**, **P. L. Blount**, **P. S. Rabinovitch**, **V. W. Rusch**, and **B. J. Reid** (1993). An endoscopic biopsy protocol can differentiate high-grade dysplasia from early adenocarcinoma in barrett's esophagus. *Gastroenterology*, **105**(1), 40–50.

199. **Li, B.** and **C. N. Dewey** (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

200. **Li, C.**, **Y. Gao**, **C. Lu**, and **M. Guo** (2023). Identification of potential biomarkers for colorectal cancer by clinical database analysis and Kaplan-Meier curves analysis. *Medicine*, **102**(6), e32877.

201. **Li, H.**, **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**, **M. Wong**, **P. J. Choi**, **L. J. K. Wee**, **A. M. Hillmer**, **I. B. Tan**, **P. Robson**, and **S. Prabhakar** (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.*, **49**(5), 708–718.

202. **Liang, J.** and **R. Liu**, Stacked denoising autoencoder and dropout together to prevent overfitting in deep neural network. *In 2015 8th international congress on image and signal processing (CISP)*. IEEE, 2015.

203. **Liaw, A.**, **M. Wiener**, *et al.* (2002). Classification and regression by randomforest. *R news*, **2**(3), 18–22.

204. **Lin, D.**, **L. Shen**, **M. Luo**, **K. Zhang**, **J. Li**, **Q. Yang**, **F. Zhu**, **D. Zhou**, **S. Zheng**, **Y. Chen**, and **J. Zhou** (2021). Circulating tumor cells: biology and clinical significance. *Signal Transduct Target Ther*, **6**(1), 404.

205. **Lin, Z.**, **G. Luo**, **W. Du**, **T. Kong**, **C. Liu**, and **Z. Liu** (2020). Recent advances in microfluidic platforms applied in cancer metastasis: Circulating tumor cells'(ctcs) isolation and tumor-on-a-chip. *Small*, **16**(9), 1903899.

206. **Liu, S.**, **B. Li**, **J. Xu**, **S. Hu**, **N. Zhan**, **H. Wang**, **C. Gao**, **J. Li**, and **X. Xu** (2020). SOD1 promotes cell proliferation and metastasis in non-small cell lung cancer via an miR-409-3p/SOD1/SETDB1 epigenetic regulatory feedforward loop. *Front Cell Dev Biol*, **8**, 213.

207. **Liu, S.**, **J. McGree**, **Z. Ge**, and **Y. Xie** (2016). Big data in healthcare applications.

208. **Liu, Y.** and **G. Zeng** (2012). Cancer and innate immune system interactions: translational potentials for cancer immunotherapy. *J. Immunother.*, **35**(4), 299–308.

209. **Lobo, I.** (2008). Chromosome abnormalities and cancer genetics.

210. **López-Larrea, C.**, **A. L. Vázquez**, and **B. S. Álvarez**, *Stem Cell Transplantation*. Springer Science & Business Media, 2012.

211. **Louis, D. N.**, **A. Perry**, **P. Wesseling**, **D. J. Brat**, **I. A. Cree**, **D. Figarella-Branger**, **C. Hawkins**, **H. K. Ng**, **S. M. Pfister**, **G. Reifenberger**, **R. Soffietti**, **A. von Deimling**, and **D. W. Ellison** (2021). The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro. Oncol.*, **23**(8), 1231–1251.

212. **Lu, Y.-T.**, **K. Delijani**, **A. Mecum**, and **A. Goldkorn** (2019). Current status of liquid biopsies for the detection and management of prostate cancer. *Cancer Manag. Res.*, **11**, 5271–5291.

213. **Ma, M.**, **H. Zhu**, **C. Zhang**, **X. Sun**, **X. Gao**, and **G. Chen** (2015). "liquid biopsy"-ctDNA detection with great potential and challenges. *Ann Transl Med*, **3**(16), 235.

214. **Ma, N.** and **S. S. Jeffrey** (2020). Deciphering cancer clues from blood. *Science*, **367**(6485), 1424–1425.

215. **Macosko, E. Z.**, **A. Basu**, **R. Satija**, **J. Nemesh**, **K. Shekhar**, **M. Goldman**, **I. Tirosh**, **A. R. Bialas**, **N. Kamitaki**, **E. M. Martersteck**, **J. J. Trombetta**, **D. A. Weitz**, **J. R. Sanes**, **A. K. Shalek**, **A. Regev**, and **S. A. McCarroll** (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**(5), 1202–1214.

216. **Madhavan, S.**, **S. Subramaniam**, **T. D. Brown**, and **J. L. Chen** (2018). Art and challenges of precision medicine: Interpreting and integrating genomic data into clinical practice. *Am Soc Clin Oncol Educ Book*, **38**, 546–553.

217. **Mafficini, A.** and **A. Scarpa** (2018). Genomic landscape of pancreatic neuroendocrine tumours: the international cancer genome consortium. *J. Endocrinol.*, **236**(3), R161–R167.

218. **Mahdizadehaghdam, S.**, **A. Panahi**, **H. Krim**, and **L. Dai** (2019). Deep dictionary learning: A PARametric NETwork approach. *IEEE Trans. Image Process.*.

219. **Maheswaran, S.** and **D. A. Haber** (2010). Circulating tumor cells: a window into cancer biology and metastasis. *Current opinion in genetics & development*, **20**(1), 96–99.

220. **Markou, A.**, **A. Strati**, **N. Malamos**, **V. Georgoulias**, and **E. S. Lianidou** (2011). Molecular characterization of circulating tumor cells in breast cancer by a liquid bead array hybridization assay. *Clin. Chem.*, **57**(3), 421–430.

221. **Marrugo-Ramírez, J.**, **M. Mir**, and **J. Samitier** (2018). Blood-Based cancer biomarkers in liquid biopsy: A promising Non-Invasive alternative to tissue biopsy. *Int. J. Mol. Sci.*, **19**(10).

222. **Martin, T. A.**, **L. Ye**, **A. J. Sanders**, **J. Lane**, and **W. G. Jiang**, *Cancer Invasion and Metastasis: Molecular and Cellular Perspective*. Landes Bioscience, 2013.

223. **Mathers, C. D.** (2020). History of global burden of disease assessment at the world health organization. *Arch. Public Health*, **78**, 77.

224. **Matsuda, S.**, **H. Takeuchi**, **H. Kawakubo**, and **Y. Kitagawa** (2017). Three-field lymph node dissection in esophageal cancer surgery. *J. Thorac. Dis.*, **9**(Suppl 8), S731–S740.

225. **McPherson, R. A.** and **M. R. Pincus**, *Henry's Clinical Diagnosis and Management by Laboratory Methods E-Book*. Elsevier Health Sciences, 2017.

226. **Mehta, S. R.**, **V. Suhag**, **M. Semwal**, and **N. Sharma** (2010). Radiotherapy: Basic concepts and recent advances. *Armed Forces Med. J. India*, **66**(2), 158–162.

227. **Mellick, A. S.**, **C. J. Day**, **S. R. Weinstein**, **L. R. Griffiths**, and **N. A. Morrison** (2002). Differential gene expression in breast cancer cell lines and stroma-tumor differences in microdissected breast cancer biopsies revealed by display array analysis. *Int. J. Cancer*, **100**(2), 172–180.

228. **Michela, B.** (2021). Liquid biopsy: A family of possible diagnostic tools. *Diagnostics (Basel)*, **11**(8).

229. **Michels, T. C.** and **K. E. Petersen** (2017). Multiple myeloma: Diagnosis and treatment. *Am. Fam. Physician*, **95**(6), 373–383.

230. **Mikolajczyk, S. D.**, **L. S. Millar**, **P. Tsinberg**, **S. M. Coutts**, **M. Zomorrodi**, **T. Pham**, **F. Z. Bischoff**, and **T. J. Pircher** (2011). Detection of EpCAM-Negative and Cytokeratin-Negative circulating tumor cells in peripheral blood. *J. Oncol.*, **2011**, 252361.

231. **Miller, M. C.**, **G. V. Doyle**, and **L. W. M. M. Terstappen** (2010). Significance of circulating tumor cells detected by the CellSearch system in patients with metastatic breast colorectal and prostate cancer. *J. Oncol.*, **2010**, 617421.

232. **Millner, L. M.**, **M. W. Linder**, and **R. Valdes, Jr** (2013). Circulating tumor cells: a review of present methods and the need to identify heterogeneous phenotypes. *Ann. Clin. Lab. Sci.*, **43**(3), 295–304.

233. **Milpied, N.**, **A. K. Fielding**, **R. M. Pearce**, **P. Ernst**, and **A. H. Goldstone** (1996). Allogeneic bone marrow transplant is not better than autologous transplant for patients with relapsed hodgkin's disease. european group for blood and bone marrow transplantation. *Journal of Clinical Oncology*, **14**(4), 1291–1296.

234. **Mina, R.**, **A. Larocca**, **M. Offidani**, **S. Bringhen**, **T. Caravita**, **V. Magarotto**, **L. Pantani**, **F. Di Raimondo**, **A. Bosi**, **I. D. Vincelli**, and **Others** (2015). Impact of complete response on survival with either autologous stem cell transplantation or conventional chemotherapy: results of a pooled analysis of 5 phase III trials in newly diagnosed multiple myeloma patients.

235. **Mirrakhimov, A. E.** (2015). Hypercalcemia of malignancy: An update on pathogenesis and management. *N. Am. J. Med. Sci.*, **7**(11), 483–493.

236. **Mohty, M.** and **J.-L. Harousseau** (2014). Treatment of autologous stem cell transplant-eligible multiple myeloma patients: ten questions and answers. *Haematologica*, **99**(3), 408–416.

237. **Morè, S.**, **L. Corvatta**, **V. M. Manieri**, **F. Saraceni**, **I. Scortechini**, **G. Mancini**, **A. Fiorentini**, **A. Olivieri**, and **M. Offidani** (2022). Autologous stem cell transplantation in multiple myeloma: Where are we and where do we want to go? *Cells*, **11**(4).

238. **Moreno, P.**, **N. Huang**, **J. R. Manning**, **S. Mohammed**, **A. Solovyev**, **K. Polanski**, **W. Bacon**, **R. Chazarra**, **C. Talavera-López**, **M. A. Doyle**, **G. Marnier**, **B. Grüning**, **H. Rasche**, **N. George**, **S. K. Fexova**, **M. Alibi**, **Z. Miao**, **Y. Perez-Riverol**, **M. Haeussler**, **A. Brazma**, **S. Teichmann**, **K. B. Meyer**, and **I. Papatheodorou** (2021). User-friendly, scalable tools and workflows for single-cell RNA-seq analysis. *Nat. Methods*, **18**(4), 327–328.

239. **Morgera, S.**, **A. K. Kraft**, **G. Siebert**, **F. C. Luft**, and **H.-H. Neumayer** (2002). Long-term outcomes in acute renal failure patients treated with continuous renal replacement therapies. *Am. J. Kidney Dis.*, **40**(2), 275–279.

240. **Murakami, F.**, **Y. Tsuboi**, **Y. Takahashi**, **Y. Horimoto**, **K. Mogushi**, **T. Ito**, **M. Emi**, **D. Matsubara**, **T. Shibata**, **M. Saito**, and **Y. Murakami** (2021). Short somatic alterations at the site of copy number variation in breast cancer. *Cancer Sci.*, **112**(1), 444–453.

241. **Mutter, R. W.**, **J. I. Choi**, **R. B. Jimenez**, **Y. M. Kirova**, **M. Fagundes**, **B. G. Haffty**, **R. A. Amos**, **J. A. Bradley**, **P. Y. Chen**, **X. Ding**, **A. M. Carr**, **L. M. Taylor**, **M. Pankuch**, **R. B. M. Vega**, **A. Y. Ho**, **P. W. Nyström**, **L. A. McGee**, **J. J. Urbanic**, **O. Cahlon**, **J. H. Maduro**, and **S. M. MacDonald** (2021). Proton therapy for breast cancer: A consensus statement from the particle therapy cooperative group breast cancer subcommittee. *Int. J. Radiat. Oncol. Biol. Phys.*, **111**(2), 337–359.

242. **Nagrath, S.**, **L. V. Sequist**, **S. Maheswaran**, **D. W. Bell**, **D. Irimia**, **L. Ulkus**, **M. R. Smith**, **E. L. Kwak**, **S. Digumarthy**, **A. Muzikansky**, **P. Ryan**, **U. J. Balis**, **R. G. Tompkins**, **D. A. Haber**, and **M. Toner** (2007). Isolation of rare circulating tumour cells in cancer patients by microchip technology. *Nature*, **450**(7173), 1235–1239.

243. **Nagy, Á.**, **G. Munkácsy**, and **B. Győrffy** (2021). Pancancer survival analysis of cancer hallmark genes. *Sci. Rep.*, **11**(1), 6047.

244. **Naithani, N.**, **A. T. Atal**, **T. V. S. V. G. K. Tilak**, **B. Vasudevan**, **P. Misra**, and **S. Sinha** (2021). Precision medicine: Uses and challenges. *Armed Forces Med. J. India*, **77**(3), 258–265.

245. **Nakano, S.**, **Y. Imawari**, **A. Mibu**, **M. Otsuka**, and **T. Oinuma** (2018). Differentiating vacuum-assisted breast biopsy from core needle biopsy: Is it necessary? *Br. J. Radiol.*, **91**(1092), 20180250.

246. **National Institutes of Health (US)** and **Biological Sciences Curriculum Study**, Understanding cancer. *In NIH Curriculum Supplement Series [Internet]*. National Institutes of Health (US), 2007.

247. **Natrajan, R.**, **A. Mackay**, **P. M. Wilkerson**, **M. B. Lambros**, **D. Wetterskog**, **M. Arnedos**, **K.-K. Shiu**, **F. C. Geyer**, **A. Langerød**, **B. Kreike**, **F. Reyal**, **H. M.**

Horlings, **M. J. van de Vijver**, **J. Palacios**, **B. Weigelt**, and **J. S. Reis-Filho** (2012). Functional characterization of the 19q12 amplicon in grade III breast cancers. *Breast Cancer Res.*, **14**(2), R53.

248. **Nesteruk, D.**, **A. Rutkowski**, **S. Fabisiewicz**, **J. Pawlak**, **J. A. Siedlecki**, and **A. Fabisiewicz** (2014). Evaluation of prognostic significance of circulating tumor cells detection in rectal cancer patients treated with preoperative radiotherapy: prospectively collected material data. *Biomed Res. Int.*, **2014**, 712827.

249. **Ni, Y.**, **F. C. Stingo**, and **V. Baladandayuthapani** (2014). Integrative bayesian network analysis of genomic data. *Cancer Inform.*, **13**(Suppl 2), 39–48.

250. **Ohuchida, K.** and **M. Hashizume** (2013). Robotic surgery for cancer. *Cancer J.*, **19**(2), 130–132.

251. **Oyajobi, B. O.** (2007). Multiple myeloma/hypercalcemia. *Arthritis Res. Ther.*, **9 Suppl 1**(Suppl 1), S4.

252. **Ozkumur, E.**, **A. M. Shah**, **J. C. Ciciliano**, **B. L. Emmink**, **D. T. Miyamoto**, **E. Brachtel**, **M. Yu**, **P.-I. Chen**, **B. Morgan**, **J. Trautwein**, **A. Kimura**, **S. Sengupta**, **S. L. Stott**, **N. M. Karabacak**, **T. A. Barber**, **J. R. Walsh**, **K. Smith**, **P. S. Spuhler**, **J. P. Sullivan**, **R. J. Lee**, **D. T. Ting**, **X. Luo**, **A. T. Shaw**, **A. Bardia**, **L. V. Sequist**, **D. N. Louis**, **S. Maheswaran**, **R. Kapur**, **D. A. Haber**, and **M. Toner** (2013). Inertial focusing for tumor antigen-dependent and -independent sorting of rare circulating tumor cells. *Sci. Transl. Med.*, **5**(179), 179ra47.

253. **O'Flaherty, L.**, **H. Wikman**, and **K. Pantel** (2017). Biology and clinical significance of circulating tumor cell subpopulations in lung cancer. *Translational Lung Cancer Research*, **6**(4), 431.

254. **Pacella, C. M.**, **G. Bizzarri**, **G. Francica**, **A. Bianchini**, **S. De Nuntis**, **S. Pacella**, **A. Crescenzi**, **S. Taccogna**, **G. Forlini**, **Z. Rossi**, **J. Osborn**, and **R. Stasi** (2005). Percutaneous laser ablation in the treatment of hepatocellular carcinoma with small tumors: analysis of factors affecting the achievement of tumor necrosis. *J. Vasc. Interv. Radiol.*, **16**(11), 1447–1457.

255. **Pacia, C. P.**, **L. Zhu**, **Y. Yang**, **Y. Yue**, **A. Nazeri**, **H. Michael Gach**, **M. R. Talcott**, **E. C. Leuthardt**, and **H. Chen** (2020). Feasibility and safety of focused ultrasound-enabled liquid biopsy in the brain of a porcine model. *Sci. Rep.*, **10**(1), 7449.

256. **Palacín-Aliana, I.**, **N. García-Romero**, **A. Asensi-Puig**, **J. Carrión-Navarro**, **V. González-Rumayor**, and **Á. Ayuso-Sacido** (2021). Clinical utility of liquid Biopsy-Based actionable mutations detected via ddPCR. *Biomedicines*, **9**(8), 906.

257. **Palumbo, A.** and **K. Anderson** (2011). Venous thromboembolism in the patient with cancer: Focus on burden of disease and benefits of thromboprophylaxis. *N. Engl. J. Med.*, **364**, 1046–1060.

258. **Pantel, K.** and **C. Alix-Panabières** (2019). Liquid biopsy and minimal residual disease - latest advances and implications for cure. *Nat. Rev. Clin. Oncol.*, **16**(7), 409–424.

259. **Pantel, K.** and **M. R. Speicher** (2016). The biology of circulating tumor cells. *Oncogene*, **35**(10), 1216–1224.

260. **Pappas, L.**, **V. A. Adalsteinsson**, and **A. R. Parikh** (2022). The emerging promise of liquid biopsies in solid tumors. *Nature Cancer*, **3**(12), 1420–1422.

261. **Park, H.-L.** and **J. Hong** (2014). Vacuum-assisted breast biopsy for breast cancer. *Gland Surg*, **3**(2), 120–127.

262. **Parker, S. H.** and **P. Z. Israel** (1996). The evolution of minimally invasive breast biopsy: from FNA to percutaneous incisional and excisional biopsy. *Surg. Technol. Int.*, **5**, 251–256.

263. **Parkin, D. M.**, **F. Bray**, **J. Ferlay**, and **P. Pisani** (2005). Global cancer statistics, 2002. *CA Cancer J. Clin.*, **55**(2), 74–108.

264. **Peiffer-Smadja, N.**, **T. M. Rawson**, **R. Ahmad**, **A. Buchard**, **P. Georgiou**, **F.-X. Lescure**, **G. Birgand**, and **A. H. Holmes** (2020). Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clinical Microbiology and Infection*, **26**(5), 584–595.

265. **Periyasamy, M.**, **H. Patel**, **C.-F. Lai**, **V. T. M. Nguyen**, **E. Nevedomskaya**, **A. Harrod**, **R. Russell**, **J. Remenyi**, **A. M. Ochocka**, **R. S. Thomas**, **F. Fuller-Pace**, **B. Győrffy**, **C. Caldas**, **N. Navaratnam**, **J. S. Carroll**, **W. Zwart**, **R. C. Coombes**,

L. Magnani, **L. Buluwela**, and **S. Ali** (2015). APOBEC3B-Mediated cytidine deamination is required for estrogen receptor action in breast cancer. *Cell Rep.*, **13**(1), 108–121.

266. **Ponziani, S.**, **G. Di Vittorio**, **G. Pitari**, **A. M. Cimini**, and **others** (2020). Antibody-drug conjugates: the new frontier of chemotherapy. *International Journal of*.

267. **Poonia, S.**, **A. Goel**, **S. Chawla**, **N. Bhattacharya**, **P. Rai**, **Y. F. Lee**, **Y. S. Yap**, **J. West**, **A. A. Bhagat**, **J. Tayal**, **A. Mehta**, **G. Ahuja**, **A. Majumdar**, **N. Ramalingam**, and **D. Sengupta** (2022). Marker-free characterization of full-length transcriptomes of single live circulating tumor cells. *Genome Res.*.

268. **Porras, T. B.**, **P. Kaur**, **A. Ring**, **N. Schechter**, and **J. E. Lang** (2018). Challenges in using liquid biopsies for gene expression profiling. *Oncotarget*, **9**(6), 7036–7053.

269. **Potdar, P.** and **N. Lotey** (2015). Role of circulating tumor cells in future diagnosis and therapy of cancer. *J. Cancer Metastasis Treat.*, **1**(2), 44.

270. **Poulet, G.**, **J. Massias**, and **V. Taly** (2019). Liquid biopsy: General concepts. *Acta Cytol.*, **63**(6), 449–455.

271. **Prince, J.** and **J. Links** (2008). Medical imaging: Signals and systems (prince, j.l. and links, j.m.; 2006) [book review]. *IEEE Signal Process. Mag.*, **25**(1), 152–153.

272. **Privitera, A. P.**, **V. Barresi**, and **D. F. Condorelli** (2021). Aberrations of chromosomes 1 and 16 in breast cancer: A framework for cooperation of transcriptionally dysregulated genes. *Cancers*, **13**(7).

273. **Probst, P.** (2020). varimp: RF variable importance for arbitrary measures. R package version 0.4.

274. **Puget, S.**, **T. Blauwblomme**, and **J. Grill** (2012). Is biopsy safe in children with newly diagnosed diffuse intrinsic pontine glioma? *Am Soc Clin Oncol Educ Book*, 629–633.

275. **Puram, S. V.**, **I. Tirosh**, **A. S. Parikh**, **A. P. Patel**, **K. Yizhak**, **S. Gillespie**, **C. Rodman**, **C. L. Luo**, **E. A. Mroz**, **K. S. Emerick**, **D. G. Deschler**, **M. A. Varvares**, **R. Mylvaganam**, **O. Rozenblatt-Rosen**, **J. W. Rocco**, **W. C. Faquin**, **D. T. Lin**, **A. Regev**, and **B. E. Bernstein** (2017). Single-Cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, **171**(7), 1611–1624.e24.

276. **Rack, B.**, **C. Schindlbeck**, **J. Jückstock**, **U. Andergassen**, **P. Hepp**, **T. Zwingers**, **T. W. P. Friedl**, **R. Lorenz**, **H. Tesch**, **P. A. Fasching**, **T. Fehm**, **A. Schneeweiss**, **W. Lichtenegger**, **M. W. Beckmann**, **K. Friese**, **K. Pantel**, **W. Janni**, and **SUCCESS Study Group** (2014). Circulating tumor cells predict survival in early average-to-high risk breast cancer patients. *J. Natl. Cancer Inst.*, **106**(5).

277. **Rafique, R.**, **S. M. R. Islam**, and **J. U. Kazi** (2021). Machine learning in the prediction of cancer therapy. *Comput. Struct. Biotechnol. J.*, **19**, 4003–4017.

278. **Ragnarsson, G.**, **G. Eiriksdottir**, **J. T. Johannsdottir**, **J. G. Jonasson**, **V. Egilsson**, and **S. Ingvarsson** (1999). Loss of heterozygosity at chromosome 1p in different solid human tumours: association with survival. *Br. J. Cancer*, **79**(9-10), 1468–1474.

279. **Rajkumar, S. V.** and **S. Kumar** (2020). Multiple myeloma current treatment algorithms. *Blood cancer journal*, **10**(9), 94.

280. **Ramalingam, N.**, **B. Fowler**, **L. Szpankowski**, **A. A. Leyrat**, **K. Hukari**, **M. T. Maung**, **W. Yorza**, **M. Norris**, **C. Cesar**, **J. Shuga**, **M. L. Gonzales**, **C. D. Sanada**, **X. Wang**, **R. Yeung**, **W. Hwang**, **J. Axsom**, **Naga Sai Gopi**, **N. D. Angeles**, **C. Greene**, **M.-F. Zhou**, **E.-S. Ong**, **C.-C. Poh**, **M. Lam**, **H. Choi**, **Z. Htoo**, **L. Lee**, **C.-S. Chin**, **Z.-W. Shen**, **C. T. Lu**, **I. Holcomb**, **A. Ooi**, **C. Stolarczyk**, **T. Shuga**, **K. J. Livak**, **M. Unger**, and **J. A. A. West** (2016). Fluidic logic used in a systems approach to enable integrated Single-Cell functional analysis.

281. **Ramirez, A. K.**, **S. N. Dankel**, **B. Rastegarpanah**, **W. Cai**, **R. Xue**, **M. Crovella**, **Y.-H. Tseng**, **C. R. Kahn**, and **S. Kasif** (2020). Single-cell transcriptional networks in differentiating preadipocytes suggest drivers associated with tissue heterogeneity. *Nat. Commun.*, **11**(1), 2117.

282. **Ranjan, B.**, **F. Schmidt**, **W. Sun**, **J. Park**, **M. A. Honardoost**, **J. Tan**, **N. Arul Rayan**, and **S. Prabhakar** (2021). scconsensus: combining supervised and unsupervised clustering for cell type identification in single-cell RNA sequencing data. *BMC Bioinformatics*, **22**(1), 186.

283. **Reid, B. J.**, **P. L. Blount**, **Z. Feng**, and **D. S. Levine** (2000). Optimizing endoscopic biopsy detection of early cancers in barrett's high-grade dysplasia. *Am. J. Gastroenterol.*, **95**(11), 3089–3096.

284. **Reynolds, H. E.** (2000). Core needle biopsy of challenging benign breast conditions: a comprehensive literature review. *AJR Am. J. Roentgenol.*, **174**(5), 1245–1250.

285. **Riethdorf, S.**, **H. Fritsche**, **V. Müller**, **T. Rau**, **C. Schindlbeck**, **B. Rack**, **W. Janni**, **C. Coith**, **K. Beck**, **F. Jänicke**, **S. Jackson**, **T. Gornet**, **M. Cristofanilli**, and **K. Pantel** (2007). Detection of circulating tumor cells in peripheral blood of patients with metastatic breast cancer: a validation study of the CellSearch system. *Clin. Cancer Res.*, **13**(3), 920–928.

286. **Ritchie, M. E.**, **B. Phipson**, **D. Wu**, **Y. Hu**, **C. W. Law**, **W. Shi**, and **G. K. Smyth** (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**(7), e47.

287. **Robinson, M. D.**, **D. J. McCarthy**, and **G. K. Smyth** (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

288. **Robinson, M. D.** and **A. Oshlack** (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, **11**(3), 1–9.

289. **Rodon, P.**, **C. Linassier**, **J. B. Gauvain**, **L. Benboubker**, **P. Goupille**, **M. Maigre**, **F. Luthier**, **J. Dugay**, **V. Lucas**, and **P. Colombat** (2001). Multiple myeloma in elderly patients: presenting features and outcome. *Eur. J. Haematol.*, **66**(1), 11–17.

290. **Rosen, E. L.**, **R. C. Bentley**, **J. A. Baker**, and **M. S. Soo** (2002). Imaging-guided core needle biopsy of papillary lesions of the breast. *AJR Am. J. Roentgenol.*, **179**(5), 1185–1192.

291. **Rosen, K.**, **V. Prasad**, and **E. Y. Chen** (2020). Censored patients in Kaplan–Meier plots of cancer drugs: An empirical analysis of data sharing. *Eur. J. Cancer*, **141**, 152–161.

292. **Roychoudhury, S.** and **S. Lahiri**, *Statistical Approaches in Oncology Clinical Development: Current Paradigm and Methodological Advancement*. CRC Press, 2018.

293. **Salahshourifar, I.**, **V. K. Vincent-Chong**, **H.-Y. Chang**, **H. L. Ser**, **A. Ramanathan**, **T. G. Kallarakkal**, **Z. A. A. Rahman**, **S. M. Ismail**, **N. Prepageran**, **W. M. W. Mustafa**, **M. T. Abraham**, **K. K. Tay**, and **R. B. Zain** (2015). Downregulation of CRNN gene and genomic instability at 1q21.3 in oral squamous cell carcinoma. *Clin. Oral Investig.*, **19**(9), 2273–2283.

294. **Salomon, R.**, **D. Kaczorowski**, **F. Valdes-Mora**, **R. E. Nordon**, **A. Neild**, **N. Farbehi**, **N. Bartonicek**, and **D. Gallego-Ortega** (2019). Droplet-based single cell RNAseq tools: a practical guide. *Lab Chip*, **19**(10), 1706–1727.

295. **Samatov, T. R.**, **A. G. Tonevitsky**, and **U. Schumacher** (2013). Epithelial-mesenchymal transition: focus on metastatic cascade, alternative splicing, non-coding RNAs and modulating compounds. *Mol. Cancer*, **12**(1), 107.

296. **Sarioglu, A. F.**, **N. Aceto**, **N. Kojic**, **M. C. Donaldson**, **M. Zeinali**, **B. Hamza**, **A. Engstrom**, **H. Zhu**, **T. K. Sundaresan**, **D. T. Miyamoto**, **X. Luo**, **A. Bardia**, **B. S. Wittner**, **S. Ramaswamy**, **T. Shioda**, **D. T. Ting**, **S. L. Stott**, **R. Kapur**, **S. Maheswaran**, **D. A. Haber**, and **M. Toner** (2015). A microfluidic device for label-free, physical capture of circulating tumor cell clusters. *Nat. Methods*, **12**(7), 685–691.

297. **Schmidts, A.**, **J. Grünewald**, **M. Kleber**, **E. Terpos**, **G. Ihorst**, **H. Reinhardt**, **G. Walz**, **R. Wäsch**, **M. Engelhardt**, and **S. Zschiedrich** (2019). GFR estimation in lenalidomide treatment of multiple myeloma patients: a prospective cohort study. *Clin. Exp. Nephrol.*, **23**(2), 199–206.

298. **Schnitt, S. J.** and **L. C. Collins**, *Biopsy Interpretation of the Breast*. Lippincott Williams & Wilkins, 2009.

299. **Schwaederle, M.**, **M. Zhao**, **J. J. Lee**, **A. M. Eggermont**, **R. L. Schilsky**, **J. Mendelsohn**, **V. Lazar**, and **R. Kurzrock** (2015). Impact of precision medicine in diverse cancers: A Meta-Analysis of phase II clinical trials. *J. Clin. Oncol.*, **33**(32), 3817–3825.

300. **Scutari, M.** (2009). Learning bayesian networks with the bnlearn R package.

301. **Seger, C.** (2018). An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing.

302. **Serrano, M. J.** and **U. Malapelle** (2023). Comments on roles of circulating tumor cells in the metastatic cascade and tumor immune escape: biology and clinical translation. *J Immunother Cancer*, **11**(1).

303. **Seyfried, T. N.** and **L. C. Huysentruyt** (2013). On the origin of cancer metastasis. *Crit. Rev. Oncog.*, **18**(1-2), 43–73.

304. **Seymour, C. W.**, **H. Gomez**, **C.-C. H. Chang**, **G. Clermont**, **J. A. Kellum**, **J. Kennedy**, **S. Yende**, and **D. C. Angus** (2017). Precision medicine for all? challenges and opportunities for a precision medicine approach to critical illness. *Crit. Care*, **21**(1), 257.

305. **Sharma, G.** and **C. Prabha** (2021). Applications of machine learning in cancer prediction and prognosis.

306. **Sharma, S.**, **A. Aggarwal**, and **T. Choudhury**, Breast cancer detection using machine learning algorithms. *In 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*. 2018.

307. **Sheikhpour, E.**, **P. Noorbakhsh**, **E. Foroughi**, **S. Farahnak**, **R. Nasiri**, and **H. Neamatzadeh** (2018). A survey on the role of interleukin-10 in breast cancer: A narrative. *Rep Biochem Mol Biol*, **7**(1), 30–37.

308. **Shenoy, A. K.** and **J. Lu** (2016). Cancer cells remodel themselves and vasculature to overcome the endothelial barrier. *Cancer Lett.*, **380**(2), 534–544.

309. **Shinozaki, A.**, Electronic medical records and machine learning in approaches to drug development. *In Artificial intelligence in Oncology drug discovery and development*. IntechOpen, 2020.

310. **Siegel, R. L.**, **K. D. Miller**, and **A. Jemal** (2015). Cancer statistics, 2015. *CA Cancer J. Clin.*, **65**(1), 5–29.

311. **Siegel, R. L.**, **K. D. Miller**, **N. S. Wagle**, and **A. Jemal** (2023). Cancer statistics, 2023. *CA Cancer J. Clin.*, **73**(1), 17–48.

312. **Simpson, E.** and **F. Dazzi** (2019). Bone marrow transplantation 1957-2019. *Front. Immunol.*, **10**, 1246.

313. **Sinha, D.**, **P. Sinha**, **R. Saha**, **S. Bandyopadhyay**, and **D. Sengupta** (2019). Improved dropclust R package with integrative analysis support for scRNA-seq data. *Bioinformatics*.

314. **Smith, D.** and **K. Yong** (2013). Multiple myeloma. *BMJ*, **346**.

315. **Solmaz, S.**, **O. Uzun**, **O. G. Sevindik**, **F. Demirkan**, **M. A. Ozcan**, **G. H. Ozsan**, and **I. Alacacıoglu** (2023). The effect of haemoglobin, albumin, lymphocyte and platelet

score on the prognosis in patients with multiple myeloma. *Int. J. Lab. Hematol.*, **45**(1), 13–19.

316. **Soper, S. A.** and **A. Rasooly** (2016). Cancer: a global concern that demands new detection technologies. *Analyst*, **141**(2), 367–370.

317. **Stott, S. L.**, **C.-H. Hsu**, **D. I. Tsukrov**, **M. Yu**, **D. T. Miyamoto**, **B. A. Waltman**, **S. M. Rothenberg**, **A. M. Shah**, **M. E. Smas**, **G. K. Korir**, **F. P. Floyd, Jr**, **A. J. Gilman**, **J. B. Lord**, **D. Winokur**, **S. Springer**, **D. Irimia**, **S. Nagrath**, **L. V. Sequist**, **R. J. Lee**, **K. J. Isselbacher**, **S. Maheswaran**, **D. A. Haber**, and **M. Toner** (2010). Isolation of circulating tumor cells using a microvortex-generating herringbone-chip. *Proc. Natl. Acad. Sci. U. S. A.*, **107**(43), 18392–18397.

318. **Stouffer, S. A.**, **E. A. Suchman**, **L. C. Devinney**, **S. A. Star**, and **R. M. Williams, Jr** (1949). The american soldier: Adjustment during army life. (studies in social psychology in world war II), vol. 1. **1**, 599.

319. **Subramanian, A.**, **P. Tamayo**, **V. K. Mootha**, **S. Mukherjee**, **B. L. Ebert**, **M. A. Gillette**, **A. Paulovich**, **S. L. Pomeroy**, **T. R. Golub**, **E. S. Lander**, and **J. P. Mesirov** (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **102**(43), 15545–15550.

320. **Sudha, B.**, **K. Suganya**, **K. Swathi**, and **others** (). Artificial intelligence is revolutionizing cancer research. *Mach. Learn. Appl. Int. J.*.

321. **Sudmant, P. H.**, **T. Rausch**, **E. J. Gardner**, **R. E. Handsaker**, **A. Abyzov**, **J. Huddleston**, **Y. Zhang**, **K. Ye**, **G. Jun**, **M. H.-Y. Fritz**, **M. K. Konkel**, **A. Malhotra**, **A. M. Stütz**, **X. Shi**, **F. P. Casale**, **J. Chen**, **F. Hormozdiari**, **G. Dayama**, **K. Chen**, **M. Malig**, **M. J. P. Chaisson**, **K. Walter**, **S. Meiers**, **S. Kashin**, **E. Garrison**, **A. Auton**, **H. Y. K. Lam**, **X. J. Mu**, **C. Alkan**, **D. Antaki**, **T. Bae**, **E. Cerveira**, **P. Chines**, **Z. Chong**, **L. Clarke**, **E. Dal**, **L. Ding**, **S. Emery**, **X. Fan**, **M. Gujral**, **F. Kahveci**, **J. M. Kidd**, **Y. Kong**, **E.-W. Lameijer**, **S. McCarthy**, **P. Flicek**, **R. A. Gibbs**, **G. Marth**, **C. E. Mason**, **A. Menelaou**, **D. M. Muzny**, **B. J. Nelson**, **A. Noor**, **N. F. Parrish**, **M. Pendleton**, **A. Quitadamo**, **B. Raeder**, **E. E. Schadt**, **M. Romanovitch**, **A. Schlattl**, **R. Sebra**, **A. A. Shabalin**, **A. Untergasser**, **J. A. Walker**, **M. Wang**, **F. Yu**, **C. Zhang**, **J. Zhang**, **X. Zheng-Bradley**, **W. Zhou**, **T. Zichner**, **J. Sebat**, **M. A.**

Batzer, **S. A. McCarroll**, **1000 Genomes Project Consortium**, **R. E. Mills**, **M. B. Gerstein**, **A. Bashir**, **O. Stegle**, **S. E. Devine**, **C. Lee**, **E. E. Eichler**, and **J. O. Korbel** (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**(7571), 75–81.

322. **Sung, H.**, **J. Ferlay**, **R. L. Siegel**, **M. Laversanne**, **I. Soerjomataram**, **A. Jemal**, and **F. Bray** (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.*.

323. **Sutton, R. T.**, **D. Pincock**, **D. C. Baumgart**, **D. C. Sadowski**, **R. N. Fedorak**, and **K. I. Kroeker** (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, **3**(1), 17.

324. **Szczerba, B. M.**, **F. Castro-Giner**, **M. Vetter**, **I. Krol**, **S. Gkountela**, **J. Landin**, **M. C. Scheidmann**, **C. Donato**, **R. Scherrer**, **J. Singer**, **C. Beisel**, **C. Kurzeder**, **V. Heinzelmann-Schwarz**, **C. Rochlitz**, **W. P. Weber**, **N. Beerenwinkel**, and **N. Aceto** (2019). Neutrophils escort circulating tumour cells to enable cell cycle progression. *Nature*, **566**(7745), 553–557.

325. **Takeshima, H.** and **T. Ushijima** (2019). Accumulation of genetic and epigenetic alterations in normal cells and cancer risk. *NPJ Precis Oncol*, **3**, 7.

326. **Tang, H.**, **H. Liu**, **W. Xiao**, and **N. Sebe** (2020). When dictionary learning meets deep learning: Deep dictionary learning and coding network for image recognition with limited data. *IEEE Trans Neural Netw Learn Syst*, **PP**.

327. **Tariyal, S.**, **A. Majumdar**, **R. Singh**, and **M. Vatsa** (2016). Deep dictionary learning. *IEEE Access*, **4**, 10096–10109.

328. **Tarver, T.** (2012). Cancer facts & figures 2012. american cancer society (ACS). *J. Consum. Health Internet*, **16**(3), 366–367.

329. **Temraz, S.**, **R. Nasr**, **D. Mukherji**, **F. Kreidieh**, and **A. Shamseddine** (2022). Liquid biopsy derived circulating tumor cells and circulating tumor DNA as novel biomarkers in hepatocellular carcinoma. *Expert Rev. Mol. Diagn.*, **22**(5), 507–518.

330. **Thery, L.**, **A. Meddis**, **L. Cabel**, **C. Proudhon**, **A. Latouche**, **J.-Y. Pierga**, and **F.-C. Bidard** (2019). Circulating tumor cells in early breast cancer. *JNCI Cancer Spectr*, **3**(2), kz026.

331. **Thiele, J.-A.**, **P. Pitule**, **J. Hicks**, and **P. Kuhn** (2019). Single-Cell analysis of circulating tumor cells. *Methods Mol. Biol.*, **1908**, 243–264.

332. **Thomas, M. C.**, **M. Brownlee**, **K. Susztak**, **K. Sharma**, **K. A. M. Jandeleit-Dahm**, **S. Zoungas**, **P. Rossing**, **P.-H. Groop**, and **M. E. Cooper** (2015). Diabetic kidney disease. *Nat Rev Dis Primers*, **1**, 15018.

333. **Tian, L.**, **X. Dong**, **S. Freytag**, **K.-A. Lê Cao**, **S. Su**, **A. JalalAbadi**, **D. Amann-Zalcenstein**, **T. S. Weber**, **A. Seidi**, **J. S. Jabbari**, **S. H. Naik**, and **M. E. Ritchie** (2019). Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods*, **16**(6), 479–487.

334. **Tickle, T.**, **I. Tirosh**, **C. Georgescu**, **M. Brown**, and **B. Haas** (2019). inferCNV of the trinity CTAT project. *Klarman Cell Observatory, Broad Institute of MIT and Harvard*.

335. **Ting, D. T.**, **B. S. Wittner**, **M. Ligorio**, **N. Vincent Jordan**, **A. M. Shah**, **D. T. Miyamoto**, **N. Aceto**, **F. Bersani**, **B. W. Brannigan**, **K. Xega**, **J. C. Ciciliano**, **H. Zhu**, **O. C. MacKenzie**, **J. Trautwein**, **K. S. Arora**, **M. Shahid**, **H. L. Ellis**, **N. Qu**, **N. Bardeesy**, **M. N. Rivera**, **V. Deshpande**, **C. R. Ferrone**, **R. Kapur**, **S. Ramaswamy**, **T. Shioda**, **M. Toner**, **S. Maheswaran**, and **D. A. Haber** (2014). Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep.*, **8**(6), 1905–1918.

336. **Tirosh, I.**, **B. Izar**, **S. M. Prakadan**, **M. H. Wadsworth**, **D. Treacy**, **J. J. Trombetta**, **A. Rotem**, **C. Rodman**, **C. Lian**, **G. Murphy**, *et al.* (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, **352**(6282), 189–196.

337. **Tošić, I.** and **P. Frossard** (2011). Dictionary learning. *IEEE Signal Process. Mag.*, **28**(2), 27–38.

338. **Tsai, W.-S.**, **J.-S. Chen**, **H.-J. Shao**, **J.-C. Wu**, **J.-M. Lai**, **S.-H. Lu**, **T.-F. Hung**, **Y.-C. Chiu**, **J.-F. You**, **P.-S. Hsieh**, **C.-Y. Yeh**, **H.-Y. Hung**, **S.-F. Chiang**, **G.-P. Lin**, **R. Tang**, and **Y.-C. Chang** (2016). Circulating tumor cell count correlates with colorectal neoplasm progression and is a prognostic marker for distant metastasis in Non-Metastatic patients. *Sci. Rep.*, **6**, 24517.

339. **Turesson, I.**, **R. Velez**, **S. Y. Kristinsson**, and **O. Landgren** (2010). Patterns of improved survival in patients with multiple myeloma in the twenty-first century: a population-based study. *J. Clin. Oncol.*, **28**(5), 830.

340. **Turner, J.** (2019). *Secondary use of electronic medical records for early identification of raised condition likelihoods in individuals: a machine learning approach*. Ph.D. thesis, City, University of London.

341. **Turner, N.**, **M. B. Lambros**, **H. M. Horlings**, **A. Pearson**, **R. Sharpe**, **R. Natrajan**, **F. C. Geyer**, **M. van Kouwenhove**, **B. Kreike**, **A. Mackay**, **A. Ashworth**, **M. J. van de Vijver**, and **J. S. Reis-Filho** (2010). Integrative molecular profiling of triple negative breast cancers identifies amplicon drivers and potential therapeutic targets. *Oncogene*, **29**(14), 2013–2023.

342. **Tuttle, K. R.**, **G. L. Bakris**, **R. W. Bilous**, **J. L. Chiang**, **I. H. de Boer**, **J. Goldstein-Fuchs**, **I. B. Hirsch**, **K. Kalantar-Zadeh**, **A. S. Narva**, **S. D. Navaneethan**, **J. J. Neumiller**, **U. D. Patel**, **R. E. Ratner**, **A. T. Whaley-Connell**, and **M. E. Molitch** (2014). Diabetic kidney disease: a report from an ADA consensus conference. *Am. J. Kidney Dis.*, **64**(4), 510–533.

343. **Urrutia, E.**, **H. Chen**, **Z. Zhou**, **N. R. Zhang**, and **Y. Jiang** (2018). Integrative pipeline for profiling DNA copy number and inferring tumor phylogeny. *Bioinformatics*, **34**(12), 2126–2128.

344. **Vaiphei, K.**, *Interpretation of Endoscopic Biopsy - Gastritis, Gastropathies and Beyond*. Springer Nature, 2021.

345. **Vamathevan, J.**, **D. Clark**, **P. Czodrowski**, **I. Dunham**, **E. Ferran**, **G. Lee**, **B. Li**, **A. Madabhushi**, **P. Shah**, **M. Spitzer**, and **S. Zhao** (2019). Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.*, **18**(6), 463–477.

346. **van Buuren, S.** and **K. Groothuis-Oudshoorn** (2010). mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.*, 1–68.

347. **van de Sant, A. J. W.**, **N. M. de Vries**, **T. J. Hoogeboom**, and **M. W. G. Nijhuis-van der Sanden** (2019). Implementation of a personalized, Cost-Effective physical therapy approach (Coach2Move) for older adults: Barriers and facilitators. *J. Geriatr. Phys. Ther.*, **42**(3), E1–E16.

348. **van Zijl, F.**, **G. Krupitza**, and **W. Mikulits** (2011). Initial steps of metastasis: cell invasion and endothelial transmigration. *Mutat. Res.*, **728**(1-2), 23–34.

349. **Varshney, K.**, **P. Ghosh**, and **A. Patel** (2022). The influence of cannabinoids on multiple myeloma cells: A scoping review. *Future Pharmacology*, **2**(3), 347–359.

350. **Vasseur, A.**, **N. Kiavue**, **F.-C. Bidard**, **J.-Y. Pierga**, and **L. Cabel** (2021). Clinical utility of circulating tumor cells: an update. *Mol. Oncol.*, **15**(6), 1647–1666.

351. **Velten, L.**, **S. F. Haas**, **S. Raffel**, **S. Blaszkiewicz**, **S. Islam**, **B. P. Hennig**, **C. Hirche**, **C. Lutz**, **E. C. Buss**, **D. Nowak**, **T. Boch**, **W.-K. Hofmann**, **A. D. Ho**, **W. Huber**, **A. Trumpp**, **M. A. G. Essers**, and **L. M. Steinmetz** (2017). Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.*, **19**(4), 271–281.

352. **Wang, H.** and **M. Song** (2011). Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming. *R J.*, **3**(2), 29.

353. **Wang, L.**, **P. Balasubramanian**, **A. P. Chen**, **S. Kummar**, **Y. A. Evrard**, and **R. J. Kinders** (2016). Promise and limits of the CellSearch platform for evaluating pharmacodynamics in circulating tumor cells. *Semin. Oncol.*, **43**(4), 464–475.

354. **Wang, S.**, **Y. Zheng**, **J. Li**, **Y. Yu**, **W. Zhang**, **M. Song**, **Z. Liu**, **Z. Min**, **H. Hu**, **Y. Jing**, **X. He**, **L. Sun**, **L. Ma**, **C. R. Esteban**, **P. Chan**, **J. Qiao**, **Q. Zhou**, **J. C. Izpisua Belmonte**, **J. Qu**, **F. Tang**, and **G.-H. Liu** (2020*a*). Single-Cell transcriptomic atlas of primate ovarian aging. *Cell*, **180**(3), 585–600.e19.

355. **Wang, S.**, **Y. Zhou**, **X. Qin**, **S. Nair**, **X. Huang**, and **Y. Liu** (2020*b*). Label-free detection of rare circulating tumor cells by image analysis and machine learning. *Sci. Rep.*, **10**(1), 12226.

356. **Wang, W.**, **V. Baladandayuthapani**, **C. C. Holmes**, and **K.-A. Do** (2013). Integrative network-based bayesian analysis of diverse genomics data. *BMC Bioinformatics*, **14 Suppl 13**(Suppl 13), S8.

357. **Wang, Y.**, **Z. Wang**, **X. Gang**, and **G. Wang** (2021). Liquid biopsy in prostate cancer: current status and future challenges of clinical application. *Aging Male*, **24**(1), 58–71.

358. **Ward, M. P.**, **L. E Kane**, **L. A Norris**, **B. M. Mohamed**, **T. Kelly**, **M. Bates**, **A. Clarke**, **N. Brady**, **C. M. Martin**, **R. D. Brooks**, **D. A. Brooks**, **S. Selemidis**,

S. Hanniffy, **E. P. Dixon**, **S. A O'Toole**, and **J. J O'Leary** (2021). Platelets, immune cells and the coagulation cascade; friend or foe of the circulating tumour cell? *Mol. Cancer*, **20**(1), 59.

359. **Warkiani, M. E.**, **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. Y. Chen**, **C. T. Lim**, and **J. Han** (2014). Slanted spiral microfluidics for the ultra-fast, label-free isolation of circulating tumor cells. *Lab Chip*, **14**(1), 128–137.

360. **Wasylewicz, A.** and **A. Scheepers-Hoeks** (2019). Clinical decision support systems. *Fundamentals of clinical data science*, 153–169.

361. **Watson, A.**, **K. Kaihara**, **J. Chew**, **A. Quinlan**, **D. Norton**, and **B. Williams** (2018). Identifying heterogeneity within rare cell populations by pairing Single-Cell RNA-Seq with cell sorting.

362. **Waxman, A. J.**, **P. J. Mink**, **S. S. Devesa**, **W. F. Anderson**, **B. M. Weiss**, **S. Y. Kristinsson**, **K. A. McGlynn**, and **O. Landgren** (2010). Racial disparities in incidence and outcome in multiple myeloma: a population-based study. *Blood*, blood–2010.

363. **Weinstein, J. N.**, **The Cancer Genome Atlas Research Network**, **E. A. Collisson**, **G. B. Mills**, **K. R. Mills Shaw**, **B. A. Ozenberger**, **K. Ellrott**, **I. Shmulevich**, **C. Sander**, and **J. M. Stuart** (2013). The cancer genome atlas Pan-Cancer analysis project.

364. **Wiedermann, C. J.** (2021). Hypoalbuminemia as surrogate and culprit of infections. *Int. J. Mol. Sci.*, **22**(9).

365. **Wiuf, C.** and **C. L. Andersen**, *Statistics and informatics in molecular cancer research*. OUP Oxford, 2009.

366. **Wolchok, J. D.**, **V. Chiarion-Sileni**, **R. Gonzalez**, **P. Rutkowski**, **J.-J. Grob**, **C. L. Cowey**, **C. D. Lao**, **J. Wagstaff**, **D. Schadendorf**, **P. F. Ferrucci**, *et al.* (2017). Overall survival with combined nivolumab and ipilimumab in advanced melanoma. *New England Journal of Medicine*, **377**(14), 1345–1356.

367. **Wolf, F. A.**, **P. Angerer**, and **F. J. Theis** (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**(1), 15.

368. **Wong, R.** and **J. Tay** (2018). Economics of multiple myeloma.

369. **Wu, F.**, **J. Fan**, **Y. He**, **A. Xiong**, **J. Yu**, **Y. Li**, **Y. Zhang**, **W. Zhao**, **F. Zhou**, **W. Li**, **J. Zhang**, **X. Zhang**, **M. Qiao**, **G. Gao**, **S. Chen**, **X. Chen**, **X. Li**, **L. Hou**, **C. Wu**, **C. Su**, **S. Ren**, **M. Odenthal**, **R. Buettner**, **N. Fang**, and **C. Zhou** (2021). Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nat. Commun.*, **12**(1), 2540.

370. **Wu, W.**, **K. Merriman**, **A. Nabaah**, **N. Seval**, **D. Seval**, **H. Lin**, **M. Wang**, **M. H. Qazilbash**, **V. Baladandayuthapani**, **D. Berry**, **R. Z. Orlowski**, **M.-H. Lee**, and **S.-C. J. Yeung** (2014). The association of diabetes and anti-diabetic medications with clinical outcomes in multiple myeloma. *Br. J. Cancer*, **111**(3), 628–636.

371. **Xiao, Y.** and **D. Yu** (2021). Tumor microenvironment as a therapeutic target in cancer. *Pharmacol. Ther.*, **221**, 107753.

372. **Xie, J.**, **R. Girshick**, and **A. Farhadi**, Unsupervised deep embedding for clustering analysis. *In* **M. F. Balcan** and **K. Q. Weinberger** (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*. PMLR, New York, New York, USA, 2016.

373. **Xu, L.**, **X. Mao**, **A. Imrali**, **F. Syed**, **K. Mutsvangwa**, **D. Berney**, **P. Cathcart**, **J. Hines**, **J. Shamash**, and **Y.-J. Lu** (2015). Optimization and evaluation of a novel size based circulating tumor cell isolation system. *PLoS One*, **10**(9), e0138032.

374. **Yang, B.**, **X. Fu**, **N. D. Sidiropoulos**, and **M. Hong**, Towards k-means-friendly spaces: Simultaneous deep learning and clustering. *In* **D. Precup** and **Y. W. Teh** (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*. PMLR, International Convention Centre, Sydney, Australia, 2017.

375. **Yang, H.-T.**, **R. H. Shah**, **D. Tegay**, and **K. Onel** (2019). Precision oncology: lessons learned and challenges for the future. *Cancer Manag. Res.*, **11**, 7525–7536.

376. **Yang, J. D.**, **M. B. Campion**, **M. C. Liu**, **R. Chaiteerakij**, **N. H. Giama**, **H. Ahmed Mohammed**, **X. Zhang**, **C. Hu**, **V. L. Campion**, **J. Jen**, **S. K. Venkatesh**, **K. C. Halling**, **B. R. Kipp**, and **L. R. Roberts** (2016). Circulating tumor cells are

associated with poor overall survival in patients with cholangiocarcinoma. *Hepatology*, **63**(1), 148–158.

377. **Yang, Y.-M.**, **T.-H. Liu**, **Y.-J. Chen**, **W.-J. Jiang**, **J.-M. Qian**, **X. Lu**, **J. Gao**, **S.-F. Wu**, **X.-T. Sang**, and **J. Chen** (2005). Chromosome 1q loss of heterozygosity frequently occurs in sporadic insulinomas and is associated with tumor malignancy. *Int. J. Cancer*, **117**(2), 234–240.

378. **Yao, X.**, **A. D. Choudhury**, **Y. J. Yamanaka**, **V. A. Adalsteinsson**, **T. M. Gierahn**, **C. A. Williamson**, **C. R. Lamb**, **M.-E. Taplin**, **M. Nakabayashi**, **M. S. Chabot**, **T. Li**, **G.-S. M. Lee**, **J. S. Boehm**, **P. W. Kantoff**, **W. C. Hahn**, **K. D. Wittrup**, and **J. C. Love** (2014). Functional analysis of single cells identifies a rare subset of circulating tumor cells with malignant traits. *Integr. Biol.*, **6**(4), 388–398.

379. **Yap, T. A.**, **D. Lorente**, **A. Omlin**, **D. Olmos**, and **J. S. de Bono** (2014). Circulating tumor cells: a multifunctional biomarker. *Clin. Cancer Res.*, **20**(10), 2553–2568.

380. **Yip, S. H.**, **P. Wang**, **J.-P. A. Kocher**, **P. C. Sham**, and **J. Wang** (2017). Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res.*, **45**(22), e179.

381. **Yu, M.**, **A. Bardia**, **N. Aceto**, **F. Bersani**, **M. W. Madden**, **M. C. Donaldson**, **R. Desai**, **H. Zhu**, **V. Comaills**, **Z. Zheng**, **B. S. Wittner**, **P. Stojanov**, **E. Brachtel**, **D. Sgroi**, **R. Kapur**, **T. Shioda**, **D. T. Ting**, **S. Ramaswamy**, **G. Getz**, **A. J. Iafrate**, **C. Benes**, **M. Toner**, **S. Maheswaran**, and **D. A. Haber** (2014). Cancer therapy. ex vivo culture of circulating breast tumor cells for individualized testing of drug susceptibility. *Science*, **345**(6193), 216–220.

382. **Zagouri, F.**, **E. Kastritis**, **A. Zomas**, **E. Terpos**, **E. Katodritou**, **A. Symeonidis**, **S. Delimpasi**, **A. Pouli**, **T. P. Vassilakopoulos**, **E. Michalis**, **S. Giannouli**, **Z. Kartasis**, **A. Christoforidou**, **K. Kokoviadou**, **E. Hatzimichael**, **D. Gika**, **C. Megalakaki**, **M. Papaioannou**, **M.-C. Kyrtsonis**, **K. Konstantopoulos**, **M. A. Dimopoulos**, and **Greek Myeloma Study Group** (2017). Hypercalcemia remains an adverse prognostic factor for newly diagnosed multiple myeloma patients in the era of novel antimyeloma therapies. *Eur. J. Haematol.*, **99**(5), 409–414.

383. **Zhang, H.**, **X. Lin**, **Y. Huang**, **M. Wang**, **C. Cen**, **S. Tang**, **M. R. Dique**, **L. Cai**, **M. A. Luis**, **J. Smollar**, **Y. Wan**, and **F. Cai** (2021). Detection methods and clinical applications of circulating tumor cells in breast cancer. *Front. Oncol.*, **11**, 652253.

384. **Zhao, W.**, **X. Li**, **J. Wang**, **C. Wang**, **Y. Jia**, **S. Yuan**, **Y. Huang**, **Y. Shi**, and **Z. Tong** (2017). Decreasing eukaryotic initiation factor 3C (EIF3C) suppresses proliferation and stimulates apoptosis in breast cancer cell lines through mammalian target of rapamycin (mTOR) pathway. *Med. Sci. Monit.*, **23**, 4182–4191.

385. **Zheng, Y.**, **D. T. Miyamoto**, **B. S. Wittner**, **J. P. Sullivan**, **N. Aceto**, **N. V. Jordan**, **M. Yu**, **N. M. Karabacak**, **V. Comaills**, **R. Morris**, **R. Desai**, **N. Desai**, **E. Emmons**, **J. D. Milner**, **R. J. Lee**, **C.-L. Wu**, **L. V. Sequist**, **W. Haas**, **D. T. Ting**, **M. Toner**, **S. Ramaswamy**, **S. Maheswaran**, and **D. A. Haber** (2017). Expression of β-globin by cancer cells promotes cell survival during blood-borne dissemination. *Nat. Commun.*, **8**, 14344.

386. **Zhou, J.** (2014). Advances and prospects in cancer immunotherapy. *New Journal of Science*, **2014**.

387. **Zhou, Y.**, **D. Yang**, **Q. Yang**, **X. Lv**, **W. Huang**, **Z. Zhou**, **Y. Wang**, **Z. Zhang**, **T. Yuan**, **X. Ding**, **L. Tang**, **J. Zhang**, **J. Yin**, **Y. Huang**, **W. Yu**, **Y. Wang**, **C. Zhou**, **Y. Su**, **A. He**, **Y. Sun**, **Z. Shen**, **B. Qian**, **W. Meng**, **J. Fei**, **Y. Yao**, **X. Pan**, **P. Chen**, and **H. Hu** (2020). Single-cell RNA landscape of intratumoral heterogeneity and immunosuppressive microenvironment in advanced osteosarcoma. *Nat. Commun.*, **11**(1), 6322.

388. **Zhou, Z.** and **M. Li** (2022). Targeted therapies for cancer. *BMC Med.*, **20**(1), 1–3.

389. **Zhu, E. Y.** and **A. J. Dupuy** (2022). Machine learning approach informs biology of cancer drug response. *BMC Bioinformatics*, **23**(1), 184.

390. **Zikos, D.** and **N. DeLellis** (2018). Cdss-rm: a clinical decision support system reference model. *BMC medical research methodology*, **18**(1), 1–14.