Leveraging Machine Learning in Identification of Biomarkers for Cancer Diagnosis and Personalized Therapy Recommendation

BY

CHITRITA GOSWAMI

(PhD17002)

Department of Computer Science

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

February 2023

Leveraging machine learning in identification of biomarkers for cancer diagnosis and personalized therapy recommendation

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

CHITRITA GOSWAMI

(PhD17002)

Department of Computer Science

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

February 2023

THESIS CERTIFICATE

This is to certify that the thesis titled **Leveraging machine learning in identification of biomarkers for cancer diagnosis and personalized therapy recommendation**, submitted by **Chitrita Goswami**, 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 him 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
Associate Professor

INDRAPRASTHA INSTITUTE OF
INFORMATION TECHNOLOGY
DELHI
IIIT Delhi, 110020

Place: New Delhi
Date: February 2023

ACKNOWLEDGEMENTS

I take this opportunity to express a deep sense of gratitude to Dr Debarka Sengupta for guiding me through the duration of my research, providing timely advice and continuous support. Dr Debarka Sengupta introduced me to the research field and helped me excel in it. He always encouraged independent thinking and motivated me to work on intellectually stimulating projects. Thanks to his encouragement, I have sustained my focus, creativity, and passion for research. He is an outstandingly supportive and understanding supervisor, and I aim to emulate his traits and outlook on life. Every time we speak, he motivates me to learn more and have a positive impact on society. I want to convey my gratitude to him for creating a pleasant and welcoming work environment that has fostered fruitful collaborations. I appreciate his unwavering support and dedication; his impact on my academic and personal growth will always be remembered. I would also like to thank Dr Lalilt Kumar of AIIMS Delhi. The collaborative experience with him was enriching and helped me see how Computer Science/AI can aid clinicians. I would also like to express my gratitude to our other collaborators at AIIMS - Dr. Deepshi Thakral and Dr. Ritu Gupta. It also taught me to understand the gap between clinical practice and theoretical solutions and how to work towards bridging that gap. I would also like to thank Dr Gaurav Ahuja, IIIT Delhi; this work would not have been possible without his valuable guidance and support. I am also grateful to everyone at the Department of Computer Science and the Department of Computational Biology. Special thanks go to the faculty members who designed excellent courses. I would like to express my gratitude towards Ms. Priti Patel and other administrative staff members who have provided valuable administrative assistance whenever required. Additionally, I would like to extend my thanks to the IT department, particularly Mr Adarsh Agarwal, for his exceptional support in resolving our technical issues. Many thanks to all my collaborators and colleagues, especially Sarita, Smriti, Neetesh,

Krishan, Shreya, Priyadarshini, Indra and Vivek, for providing a healthy research environment to conduct scientific work. I would also like to thank my mates from other labs in the institute who helped me expand my horizon and made research life fun - Aanchal Mongia, Rahul Gangopadhyay, Anjali Lathwal, Garima Budhani, Anupriya, Praveen, Sagnik, Ramneek and so many other wonderful souls who enriched me both with their friendship and their knowledge. I would also take this opportunity to specially appreciate the role of some of my great friends, Pushpita, Namrata, Ayan, Mitali, Aakarsh, Soumyadeep, Koustov, Venki and Tharma. They stood as pillars of support during this journey. Their constant encouragement made it infinitely easier to deal with the ups and downs of research life. I would also like to express my thanks and gratitude to all the undergrads and MTech students I got the opportunity to mentor. The experiences honed my skills as a mentor and as a collaborator and also expanded my knowledge. I would also like to thank IIIT Delhi for funding my Doctoral Research.

Last but not least, I thank my parents for their unwavering support throughout. Without their support, none of my endeavors would be successful. I would specially like to thank my father for always instilling in me that I could do whatever I wanted to do and be whomever I wanted to be.

Chitrita Goswami

ABSTRACT

In an era where machine learning (ML) is changing the landscape of financial markets, education, security and privacy, the retail sector, and many other crucial aspects of human life, it is only fitting that we should use its potential for personalized medicine. Combining precision medicine with statistical analysis and machine learning techniques may pave the future of disease treatment. Personalized, or precision, medicine consists of using knowledge specific to a patient, such as biomarkers, genomic information, demographics, or lifestyle characteristics, best to treat their ailment, rather than generic best practices. According to a given scenario, machine learning (ML) can help predict the best treatment plan for the patient. ML can help supply clinicians with high-confidence hypotheses to support the complex decision-making process on an individual basis. This system of assistance is called clinical decision support systems (CDSS). Because cancer is so heterogeneous in nature, it is essential that each patient's treatment be individually tailored and targeted rather than adopting a standard system. Some key aspects of clinical decision-making are improving treatment efficiency, reducing adverse effects, lowering patient and care providers' costs, and diagnosing the disease early.

To study, design, analyze and interpret such multidisciplinary aspects of clinical and translational cancer research, we drew on both statistical and machine-learning based methods. Below is an anecdote of our key contributions that successfully incorporate machine learning, genomics and patient-focused healthcare.

We start the journey at the cellular level, where we propose a method that can help reveal the factors contributing to cellular heterogeneity in single-cell datasets. By identifying influential genes that contribute to cellular heterogeneity, our proposed method *InGene* lays the groundwork for personalized medicine.

Single-cell RNA sequencing (scRNA-seq) provides a powerful means of characterizing transcriptional heterogeneity within cells of seemingly identical phenotypes. Due to factors like high variability in scRNA-seq data, high dimensionality, and sparsity, traditional feature selection methods fall short in this task. Recently, non-linear dimensionality reduction techniques have made foray into scRNA-seq as they help us assess local and global cellular arrangement.

However, non-linear dimensionality reduction techniques are primarily used for visualization purposes only since they do not shed any light on the individual genes' identity that influences the non-linear transformation. We developed *InGene*, a first of its kind non-linear unsupervised method to overcome this limitation. Our method can also be used as an alternative to state-of-the-art methods for finding differential genes, which can be further used as a targeted sequencing panel, thus aiding in clinical decision making. *InGene* can be used to obtain reliable targeted panels for scRNA-sequencing, thus reducing the cost manifold. Using a cost-effective scRNA-seq sequencing solution can prove to be a headway in personalized therapy recommendation and help make the clinical decision making process more effective.

Next, we expand the scope from cellular insights to a broader patient-centric approach. In the realm of oncology, there is a critical need for diagnostic methodologies that are both efficacious and patient-friendly. In this chapter, we contribute to improving cancer diagnosis. Our study proposes an affordable, non-invasive, liquid-biopsy based diagnostic method. Although tissue biopsy is widely used to diagnose cancer, it has drawbacks, particularly when repeated sampling is necessary. Due to their ability to precisely identify the existence and subtype of tumours, tumour educated platelets (TEPs) have recently attracted interest. The majority of research involving TEPs has utilized marker-panels that include hundreds of genes, which can be expensive and impede the adoption of the diagnostic method. To address this issue, we investigated TEP expression profiles that are available to the public and discovered a signature of 11 platelet-genes that can effectively differentiate between malignant and normal samples.

Next, in our journey, we foray from disease detection to disease management. We

propose to enhance patient outcomes for Multiple Myeloma (MM) patients and help clinicians optimize a patient's treatment plans. Patient stratification and prediction of disease recurrence is another important aspect of personalized therapy. To determine the probability of recurrence in MM patients receiving Autologous Stem Cell Transplantation (ASCT), we developed a stratification model to enhance prognosis estimate and treatment efficacy. For a lot of practical reasons, it is crucial to identify whether a patient undergoing ASCT is at high risk for recurrence (likely to relapse within 36 months). Our model, which consists of a 3-factor multivariate 2-stage staging system, is highly decisive in predicting the outcome of stem cell rescue. It is essential to detect cancer promptly in order to manage cancer patients effectively.

In conclusion, this thesis harmonizes molecular insights, diagnostic innovations and clinical management in oncology.

LIST OF PUBLICATIONS AND PATENT

PUBLICATIONS

1. Chitrita Goswami, Sarita Poonia, Lalit Kumar, Debarka Sengupta

   Staging System to Predict the Risk of Relapse in Multiple Myeloma Patients Undergoing Autologous Stem Cell Transplantation

   Frontiers in Oncology, Volume 9, Article 633, (2019).

2. Chitrita Goswami, Smriti Chawla, Deepshi Thakral, Himanshu Pant, Pramod Verma, Prabhat Singh Malik, Jayadeva, Ritu Gupta, Gaurav Ahuja, Debarka Sengupta

   Molecular signature comprising 11 platelet-genes enables accurate blood-based diagnosis of NSCLC

   BMC Genomics, Volume 21, Article 744, (2020).

3. Chitrita Goswami, Debarka Sengupta

   InGene: Finding influential genes from embeddings of nonlinear dimension reduction techniques

   bioRxiv, doi: https://doi.org/10.1101/2023.06.19.545592, (2023).

PATENT

1. Sengupta D, **Goswami C**, Chawla S.

   Gene panel for detecting the presence of blood-based genetic markers of non-small cell lung cancer (202011042049: Provisional patent filed on 28.09.2020. IIIT-D transferred the technology (valued at Rs. 40 Lakhs) to CareOnco Biotech Pvt. Ltd.)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

Abbreviations	Meaning
NSCLC	Non-Small Cell Lung Cancer
PCA	Principal Component Analysis
scRNA-seq	Single-cell RNA sequencing
ML	Machine Learning
CDSS	Clinical Decision Support Systems
DSS	Durie Salmon Staging
ISS	International Staging Systems
UV	Ultraviolet
NGS	Next-generation high throughput sequencing
EHR	Electronic Health Records
TEP	Tumour Educated Platelets
ctDNA	Circulating Tumour DNA
cfDNA	Circulating free DNA
RNA-seq	RNA sequencing
RNA	Ribonucleic Acid
AUC	Area Under Curve
RF	Random Forest
LDA	Linear Discriminanat Analysis
RT-qPCR	Real Time qPCR
qPCR	Quantitative Polymerase Chain Reaction
GB	Gradient Boosting
TF	Transcription Factors

CHAPTER 1

INTRODUCTION

1.1 How cancer starts

To understand cancer and its evolution, let us start with a cell. The smallest component of a living thing is a cell. The DNA in human cells contains the distinctive genetic code. The nucleus, a component of the cell contains majority of the DNA [1]. Proteins are necessary for growth and functioning of the body. Proteins are encoded by genes, a tiny fragment of DNA. Each gene must possess the correct set of instructions for the synthesis of proteins [2]. This enables the protein to carry out the proper task for the cell. All of the genetic data in the DNA must be copied by the replicating cells in order for the new cells to have everything they need to operate properly. Sometimes mistakes can happen during the replication of DNA, leading to mutations. One must note that mutations in DNA can arise from a variety of mechanisms beyond errors during replication, such as - chemical mutagens, physical mutagens such as Ultraviolet(UV) radiation, DNA strand slippage etc. As a result of these mutations the sequence of the smaller molecules that make up a gene is frequently altered. A broken gene may result in the creation of a faulty protein, which may result in aberrant cellular changes. As a result of these faults not being repaired, they are passed down from parent cells to their progeny and eventually add up with each subsequent generation of cells. The ensuing dysfunctional proteins may alter the rates of cell division or obstruct the system's usual controls, such as cell cycle arrest or programmed cell death, disrupting the normal operations of the cell [3]. This in turn may cause cells to multiply uncontrollably. Such unchecked cell growth and the spread of abnormal cells leads to cancer. In addition to uncontrolled growth, cancer cells also fail to communicate with other cells, which results in the loss of normal restraints on growth. Cancer cells can also invade and destroy adjacent tissues and spread to other locations [4].

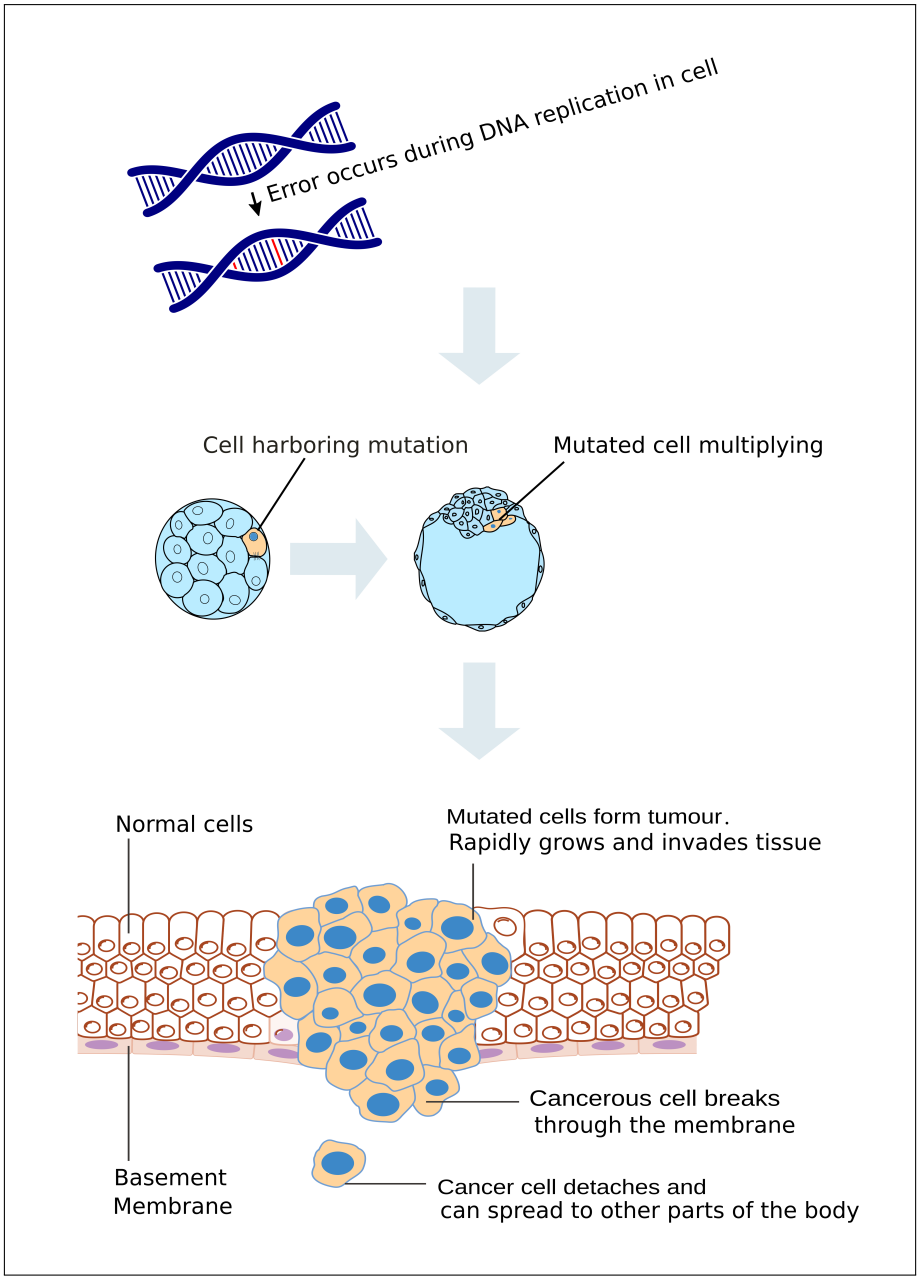

Figure 1.1: **Evolution from error in DNA replication to tumours.**

To summarize, cancer is the condition that develops when cells in a particular area of the body begin to proliferate uncontrollably. Tumours are formed by these cells and are lumps or masses. Cancer as a disease of multicellularity arises from the concept that the uncontrolled growth of cells is only problematic in a multicellular context. In single-celled organisms, proliferative advantages could simply enhance the fitness of that particular cell. But in multicellular organisms, where cellular cooperation is crucial for the functioning of the organism as a whole, uncontrolled cellular growth disrupts this cooperative structure [5]. The primary cause of cancer are DNA mutations.

Cells may begin to grow uncontrollably and may lead to the development of tumours when mutations occur. One should note that the term "cancer" doesn't refer to one specific illness, but rather to a wide range of disorders that comprise of uncontrolled cell development. Tumours belong to mainly two categories: benign and malignant. Benign tumours tend to develop gradually and do not metastasize (spread to other regions of the body). Thus, they are simpler to treat and can be eliminated quickly. On the other hand, malignant tumours do tend to spread [4]. The term "metastasis" refers to this mechanism. Because they can spread to other sites and develop new tumours, metastatic malignancies are more deadly than cancers that only affect the main site. These tumours can encroach on neighboring organs and tissues and spread to different areas of the body via the lymphatic or blood systems. There are more than 100 distinct varieties of malignant growth since it can develop practically everywhere on the body. Even though cancer is an immensely complex and diverse disease; a set of characteristics are shared among almost all malignancies. Those characteristics, named hallmarks of cancer, are a unified set of capabilities that are acquired during tumourgenesis [6]. Tumourgenesis is a multi-step process.

The first phase of this process is the **initiation** phase. Usually, when a cell loses contact with the organism that contains it, it will receive a signal to apoptose or die. This is called contact inhibition, and it prevents excessive cell growth. However, in the initiation phase of tumour growth, this signal is blocked. This allows the cell to begin growing out of control. By interrupting these signals that promote growth, cancer cells are able to take charge of their own cell fate [6].

The next step in the process is called the **promotion** phase. The growth of a cell involves signals that tell it to grow and divide, which the cell does. However, these signals get blocked in the promotion phase of tumour growth. Thus, a cell doesn't die, even when it grows too large [6].

The final step in the process is the **progression** phase. It involves the growth of new blood vessels or, neovascularization [6]. The nutrition and oxygen provided by the new blood vessels that the tumour produces are essential for it's development and survival.

Angiogenesis is multi-step, complex process that results in the growth of new blood vessels. It is regulated by a balance between several elements that either promote or impede blood vessel growth. Tumours trigger something called an angiogenic switch that upsets this equilibrium. This imbalance can lead to the formation of new blood vessels in an unregulated manner, which in turn can increase the malignant properties of the tumour [6].

The hallmarks of cancer are not limited to specific types of cells; they can occur in any cell in the body. In summary, cancer is a disease that occurs when cells become independent from the organism they originated from.

1.2 Technological advancements are helping us bring solutions from bench to bedside

The diagnosis, treatment, and care for any disease involve intricate steps and thus is, a complex process. We should also take into account that it's not always possible to predict a patient's response to a treatment routine with objective certainty. There are several factors that may play a part in a patient's response to the treatment procedure, such as age, gender, genetics, comorbidities, and response to previous therapies. Thus, it is desirable that a treatment should be tailored to the individual patient's needs. A complex disease like cancer makes the process even more challenging. Therapies designed in the laboratory often fail in real-world scenarios because laboratory experiments are conducted in controlled environments with many constraints in place. The gap is significant, and technology has played an enormous role in helping bridge that gap. Any technological advancement - from Radiography or X-ray imaging to Magnetic Resonance Imaging (MRI) has led to a better understanding of the human body and, thus, diseases. Technological advancements in the medical field have enabled doctors to make more accurate diagnoses and have generated a wealth of data that has dramatically enhanced our understanding of diseases. Cancer detection and treatment have also greatly benefited from such advancements.

1.2.1 Previous work

Till the early 19th century, cancer detection was primarily based on physical examination and observation of symptoms, and surgery was the primary form of treatment. Then, scientific oncology emerged, using the modern microscope to study diseased tissues. Rudolf Virchow, considered the founder of cellular pathology, connected microscopic pathology to illness and provided a scientific basis for the modern pathologic study of cancer. This approach improved the understanding of the damage caused by cancer and helped the development of cancer surgery [7].

Cancer imaging began with X-rays in the early 1900s [8]. It enabled clinicians to identify tumours that could not be detected by physical inspection, alone. The early 20th century saw the introduction of chemotherapy. Chemotherapy, incorporated the use of chemicals to treat cancer. The invention of chemotherapy was a significant step forward in treating cancer that had spread [9]. In the second half of the 1900s, radiation therapy improved with new methods like IMRT and stereotactic radiosurgery, which target tumours with high radiation doses while reducing harm to healthy tissue. The use of computers in the same period greatly changed cancer detection and treatment. Digital imaging technologies like CT and MRI were introduced, producing detailed body images that could be analyzed with computer software [10]. This has greatly improved the accuracy of cancer detection and staging.

The rise in computer use and accessibility led to the introduction of Electronic Health Records (EHRs). EHRs give healthcare providers a complete picture of a patient's medical history, including past diagnoses, treatments, and results [11]. In cancer care, this information helps providers create personalized treatment plans for patients, improving outcomes and reducing the chance of cancer returning. EHR and tumour registry data was successfully used by [12] to efficiently identify individuals with a previous cancer history in a health plan. Such insights can help improve treatment efficacy and reduce the economic burden for patients and care providers, among other benefits.

In the early 2000s, Next-generation high throughput sequencing (NGS) technologies came into the foray. It transformed the landscape of disease detection and treat-

ment. NGS technologies became a widely adopted tool in cancer research as they allow to uncover genomic, transcriptomic, and epigenomic scenes of individual malignant growths [13]. NGS also enables the analysis of multiple genes cost-effectively and has been applied in examining clinical cancer samples and offering NGS-based molecular diagnosis.

1.2.2 Current Landscape

Traditionally cancer detection relied on imaging and invasive biopsies. Also, cancer treatment has historically been determined by a tumour's tissue of origin [14]. Today, empowered by the latest technological advances, scientists can look further. The last decade ushered in the dawn of big data. The explosion of data sources and the availability of faster and more powerful computational units led to an era of data-driven decision-making [15]. Machine learning, a subset of artificial intelligence, makes use of statistics and pattern recognition algorithms to identify patterns and relationships that may not be immediately apparent and make predictions or decisions based on these patterns.

Thus biomedical researchers started to take advantage of the data boom and use machine learning on biomedical data to derive insights and make predictions [15]. Research is being carried out on using machine learning in identifying biomarkers for cancer diagnosis and personalized medicine [16, 17]. A recent article by News Medical Life Sciences discusses how AI and machine learning could improve cancer diagnosis through biomarker discovery [18]. The article mentions that current interest in biomarkers spans the need for personalized cancer therapy and monitoring for disease progression and recurrence to cancer risk assessment and early detection.

We can use digitized clinical data to hypothesize about the patient's survival [19]. This can help influence the next course of action in the treatment plan. Finding patterns in clinical data may also improve prognosis estimation, the possible reaction of the patient to a certain drug, etc. One example of this can be seen with the Tamoxifen drug. Patients diagnosed with estrogen receptor (ER) positive breast cancer often receive Ta-

moxifen as part of their therapeutic regimen. By analyzing clinical data and genomic information, studies have observed patterns correlating certain gene expressions with the efficacy of Tamoxifen. For instance, a lower expression of the gene CYP2D6 has been associated with reduced metabolism of Tamoxifen, leading to lower therapeutic efficacy. Therefore, patients exhibiting this genomic pattern might require alternative treatment strategies or dose adjustments. By integrating this genomic data with clinical patient data, clinicians can make more informed decisions regarding the potential success of Tamoxifen therapy for individual patients [20]. The amalgamation of technology and pattern recognition has yielded not only the identification of novel biomarkers but also a novel class of biomarkers. Tissue biopsy, apart from being invasive, just offers a one-time snapshot of the disease life-cycle, obscuring the leads for potential course corrections. Liquid biopsy technology has emerged as a promising alternative aimed at overcoming these limitations. Blood-based biomarkers, such as circulating tumour cells and cell-free DNA, can provide information about the presence and progression of cancer. Machine learning can be leveraged to analyze the data and discover new signatures [21, 22].

Recent developments in high-throughput sequencing have enabled us to profile transcriptomes even at the single-cell level. This has led to a greater understanding of cell populations' heterogeneity, disease states, identifying cell types, developmental lineages, etc. Single cell sequencing has applications beyond the actual tumour cells. Increasing evidence shows that non-malignant cells within and around tumours have essential roles in disease progression. Thus, with the advent of single cell sequencing, scientists are now afforded a more powerful lens and can look at sub-populations of cells. Single-cell data can provide information about the genetic and molecular characteristics of individual cells, which can be used to identify specific cancer subtypes and predict treatment responses. Machine learning algorithms are being used extensively to cluster and classify cell types more accurately, leading to more reliable downstream analysis results [23, 24].

These techniques provide a more comprehensive and multi-modal picture (Figure 1.2). Together, these factors have led to the popularity of machine learning for

biomarker discovery and personalized medicine in the last decade. The future of precision oncology is significantly tied to such new technologies that allow us better to understand the heterogeneity of tumours at a molecular level.

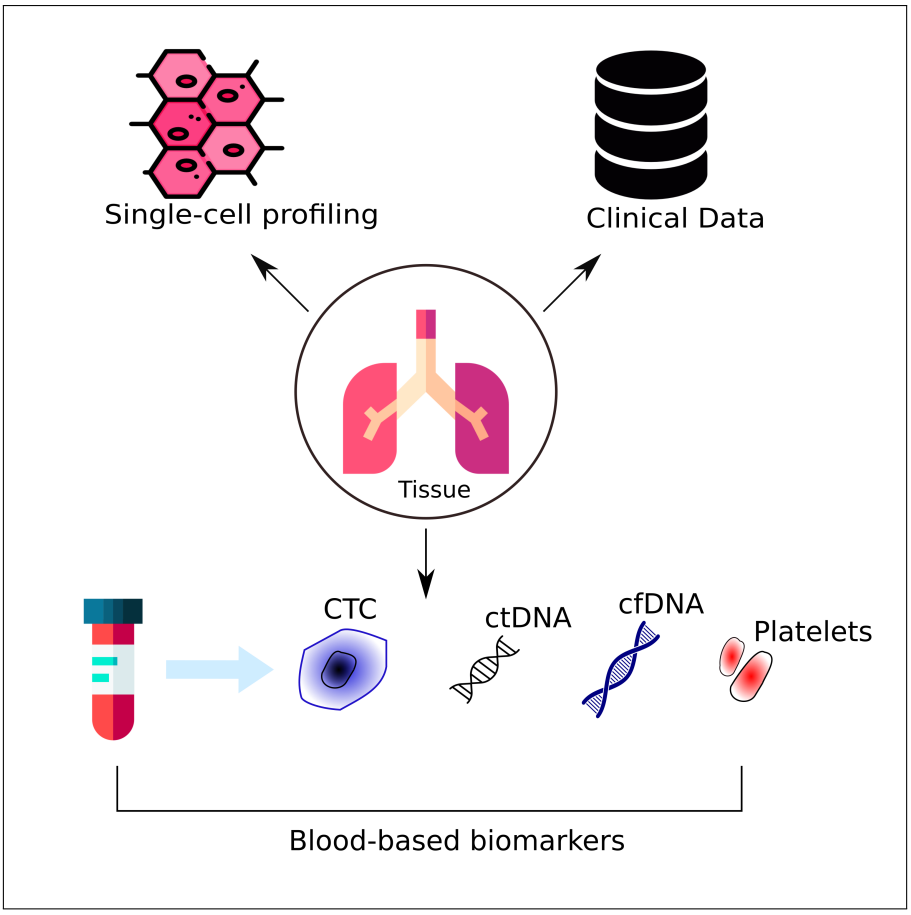

Figure 1.2: **Multi-modal biomarkers: Digitized clinical data, blood-based biomarkers and single-cell profiling contribute in development of affordable and personalised medicine in cancer.**

1.3 Need for affordable molecular diagnostics and personalised medicine in cancer

1.3.1 Economic burden of cancer

Cancer results in economic burden for patients, healthcare systems, and countries due to healthcare spending, and productivity losses from morbidity and premature mortality.

In 2017, estimated cancer healthcare spending was USD 161.2 billion in USA. USD

150.7 billion was lost due to premature mortality. The European Union reports state that healthcare spending for cancer was €57.3 billion, and losses due to premature death was €47.9 billion [25]. A study [26] used two-part models and found that cancer patients had nearly 4 times higher mean expenditures per person than those without cancer.

The cost of cancer care is not distributed evenly throughout the population. The impact is disproportionately borne by individuals with lower incomes and those without health insurance. In developing and underdeveloped nations, an even greater economic burden is posed on patients and their families. The financial burden associated with cancer treatment can force patients and households to acute misery and even insolvency [27]. Access to screening and diagnostic facilities bear additional costs for rural populations, as often such centres exist only in a few cities. It may be completely infeasible for a section of the populace in such countries to consider certain types of organized screening if no insurance system is in place to finance the screening costs, much less the treatment of the cases diagnosed. [28]

Trends signify that we can expect greater intensity of health care service use and increasing costs of cancer care in the future [29, 30, 31]. This shall result in a greater burden of cancer. Therefore, affordable screening and diagnostics are need of the hour. Designing affordable screening technologies shall also encourage people to get routine checkups, and possibly lead to timely detection of the disease. Early diagnosis is crucial for effective medical management of cancer patients. Early diagnosis contributes in the improvement of patient survival, reducing treatment cost and thus also helps to reducing the economic burden.

1.3.2 Designing tailored therapies

Cancer is a heterogeneous disease. Heterogeneity in genomic and phenotypic features has been observed not only between individual patients but also in individual tumour regions [32]. This comprehensive review by McGranahan and Swanton provides a detailed analysis of intratumoral heterogeneity, highlighting the implications of this heterogeneity for tumor evolution, treatment resistance, and therapeutic strategies. The

authors discuss the presence of distinct subclonal populations within individual tumors and the challenges they pose for precision oncology. The recent advent of genomic profiling and the increasing affordability of sequencing have revealed previously unknown characteristics of tumour heterogeneity. This inspired changes in clinical decision making for oncology treatment and management. [33].

Every patient's response to cancer treatment is different. Predicting response and, therefore, survival is essential for the clinician and patient to make informed decisions. A patient may choose to forego an expensive or invasive therapy if their predicted chance of survival is low. In such a case, a clinician would be encouraged to look for better suited alternatives for disease management. Personalized medicine represents an approach to patient care that allows doctors to select treatments most likely to help patients based on a genetic understanding of their disease. Tailored therapies in oncology, therefore, leverage the unique genetic and molecular profile of a patient's tumour to deliver treatment specifically effective against that tumour. This personalized approach has several benefits, including increased effectiveness, reduced side effects, and improved patient outcomes [34].

The era of targeted therapy was kick-started with the development of imatinib (Gleevec) for treating chronic myeloid leukaemia (CML). It works by specifically inhibiting the BCR-ABL fusion protein that is characteristic of CML [35]. This precision attack led to dramatically improved patient outcomes and set the stage for more tailored therapies.

HER2, a protein, promotes the growth of cancer cells. The HER2 gene has been found to be over-expressed in 20-25 percent of breast cancer cases. This leads to aggressive tumour growth [36]. Trastuzumab, or Herceptin, is a monoclonal antibody developed to target HER2 positive breast cancers specifically. It significantly improved patient survival rates [37]. This is another example of the use of personalized therapies targeting HER2.

In addition, the development of immune checkpoint inhibitors like pembrolizumab (Keytruda), which targets PD-1 and is used in treating several cancers, including

melanoma and non-small cell lung cancer, demonstrates the effectiveness of tailored therapy. It works by enabling T-cells to recognize and attack cancer cells [38]. Thus, there has been an increase in evidence suggesting that genomics predicts cancer prognosis. The ambiguous nature of the data concerning the relationship between genetic profiles and prognosis has significantly fueled the need for personalized medicine [39]. Therefore, in the past decade, we have seen a shift from a "one size fits all" approach to a more tailored therapy for cancer.

1.4 Overview of popular sequencing methods and their analysis

1.4.1 RNA sequencing

RNA sequencing, also known as RNA-Seq, is a method that uses high-throughput sequencing techniques to analyze the transcriptome of a cell. RNA-Seq provides greater coverage and more precise information about the dynamic nature of the transcriptome than previous methods, such as Sanger sequencing and microarrays [40]. In addition to measuring gene expression levels, RNA-Seq data can reveal novel transcripts, detect alternatively spliced genes, and identify allele-specific expression [41]. Recent improvements in RNA-Seq methods, from sample preparation to data analysis, have made it possible for us to gain a deeper understanding of transcriptomic complexity. RNA-Seq can be used to analyze different types of RNA, including mRNA, pre-mRNA, total RNA, and non-coding RNA, such as microRNA and long non-coding RNA. Since the discovery of RNA's crucial role as an intermediate between the genome and proteome, identifying transcripts and measuring gene expression has been a fundamental part of molecular biology. This process involves extracting RNA, enriching for mRNA, synthesizing cDNA, preparing a sequencing library with adaptors, and sequencing the library using a high-throughput platform [40].

Computational methods are then used to align or assemble the sequencing reads to

the transcriptome and quantify gene expression. Bulk RNA-seq experiments can estimate the total gene expression levels of a heterogeneous population of cells. Technological advances in wet-lab and computational domains have greatly improved RNA-seq, providing an unbiased and more transparent view of RNA biology and transcriptomes. Bulk RNA-seq is widely used to enhance our understanding of cancer biology and has the potential to develop new clinical applications [42]. However, using mean gene expression from bulk RNA-seq profiles can obscure the true signals from subpopulations of cell types that may be driving biological processes. This problem has led to the development of new single-cell RNA sequencing (scRNA-seq) technology [43].

1.4.2 Single-cell RNA sequencing

Single-cell RNA sequencing (scRNA-seq) is a technique that measures the gene expression of individual cells, rather than averaging the expression of a population of cells. This allows scientists to discover the diversity and complexity of cell types, functions and interactions within tissues, organs and organisms [43]. Single-cell measurement techniques such as flow cytometry and fluorescence in situ hybridization (FISH) have been around for a long time and are commonly used in labs. However, these methods have limited use as they can only profile a small number of genes and proteins, which limits the amount of information obtained from single-cell samples. Recently, new technologies have emerged in single-cell isolation and profiling of genomics, transcriptomics and proteomics, which have enabled single-cell analysis to scale up [43]. Single-cell sequencing technologies that aim to profile RNA transcriptomes encounter difficulties when quantifying different RNA species. To address this challenge, several methods have been developed to amplify small amounts of mRNA within a single cell. A recent technique called Drop-Seq has gained immense popularity, due to its advantage in scalability. Drop-seq is a microfluidic-based method that isolates individual cells in droplets and then barcodes and amplifies their mRNA transcripts [44]. Drop-Seq uses droplets to compartmentalize cells into nanoliter-sized reaction chambers for analysis of their mRNA transcripts while remembering the transcripts' cell of origin, using a

molecular barcoding strategy. This method uses a microfluidic device to compartmentalize droplets containing a single cell, lysis buffer, and a microbead covered with barcoded primers. With this technique, one single scientist can make approximately 10000 single-cell libraries every day in parallel inexpensive and easy experiments. Another technique is inDrop [45], which is used for high-throughput single-cell labeling. This approach is similar to Drop-seq but uses hydrogel microspheres to introduce oligonucleotides. A study comparing six prominent scRNA-seq methods: CEL-seq2, Drop-seq, MARS-seq, SCRB-seq, Smart-seq, and Smart-seq2 [46] found that while Smart-seq2 detected the most genes per cell and across cells, CEL-seq2, Drop-seq, MARS-seq, and SCRB-seq quantified mRNA levels with less amplification noise due to the use of unique molecular identifiers (UMIs).

In parallel to these developments, single nuclei sequencing (snRNA-seq) has emerged as a pivotal method, especially useful for analyzing tissues that are difficult to dissociate or contain a high proportion of non-viable cells [47]. Like scRNA-seq, snRNA-seq allows for the examination of gene expression at an individual cell level. However, instead of whole cells, snRNA-seq isolates and sequences RNA from individual nuclei [48]. This is particularly advantageous when dealing with complex tissues such as the brain, where cell density and fragility can pose challenges to traditional single-cell sequencing methods. The snRNA-seq method involves isolating nuclei, usually through a gentle lysis process that preserves the nuclear RNA, followed by similar barcoding and sequencing steps as found in scRNA-seq [49]. This approach allows researchers to study not only the transcriptome but also aspects of chromatin organization and regulatory sequences that are retained within the nucleus. While such methods, including single nuclei sequencing, have greatly improved the ability to amplify small quantities of mRNA in single cells or nuclei for scRNA-seq and snRNA-seq, they still have limitations and challenges that can affect their accuracy and sensitivity. These include:

- Amplification bias: Some methods can introduce bias in the amplification process, leading to inconsistent representation of certain genes in the final sequencing data.
- Drop-out events: Due to the small amount of starting material, some genes may

not be detected or may have low coverage in single-cell sequencing data, leading to "drop-out" events.

- Technical noise: Technical noise can arise during the amplification and sequencing steps, resulting in false positive or negative signals.
- High costs: Some methods can be costly and time-consuming, limiting their application for large-scale studies.
- Limited cell throughput: Some methods are limited in their ability to process a large number of cells, which can be problematic for studies requiring high cell throughput.
- Compatibility issues: Some methods may not be compatible with certain types of samples, such as degraded or low-quality RNA.

Despite these limitations, scRNA-seq and snRNA-seq are the most comprehensive technology that can reveal how genes are regulated in different cell states and stages of development. RNA-seq assumes that all cells in a sample have similar gene expression profiles, which may not be accurate for heterogeneous tissues or complex biological systems. scRNA-seq can capture the variability and heterogeneity of cell populations and identify rare or novel cell types. RNA-seq cannot track the temporal dynamics or lineage relationships of cells during development or differentiation. scRNA-seq can reconstruct the trajectories and transitions of cells along developmental pathways or cellular fates. Therefore, scRNA-seq fills the gap of RNA-seq by providing a more detailed and comprehensive view of gene expression at the single-cell level.

1.5 Computational challenges

1.5.1 Handling clinical data

In the last decade healthcare data has become increasingly available, as more and more hospitals and clinicians are adopting electronic record keeping. This has led to an increased potential for knowledge discovery and thus, personalized medicine.

Electronic Health Records (EHRs) are proving to be a very valuable resource. It provides real-time patient-centred records that can be accessed securely and instantly by

authorized users. However, integrating clinical data presents significant computational challenges [11].

One of the primary challenges in analyzing clinical data is dealing with the unstructured nature of the data. EHRs contain a mix of structured and unstructured data, including free text notes, imaging results, and laboratory values. This mix of data types can make it challenging to analyze and develop models using clinical datasets [50].

Furthermore clinical datasets often consist of a limited number of samples which is often a bottleneck for machine learning and deep learning algorithms. Missing data is another common problem in clinical datasets. It can occur due to patients dropping out of a study or absent information for a given variable [51].

The high degree of variability between patients is another crucial challenge which can make it difficult to identify patterns, especially when dealing with rare and heterogeneous diseases. To overcome these challenges, algorithms must be designed to take into account the unique characteristics of clinical datasets, including the need to handle missing data, account for patient variability, and identify meaningful associations despite the small sample sizes.

Overall, while the integration of clinical data presents many challenges, the potential for knowledge discovery and personalized medicine makes it an important area of research. By developing new computational approaches that can handle the unique characteristics of clinical datasets, we can unlock the full potential of EHRs and improve patient outcomes.

1.5.2 Genomics data analysis

Small sample size is a common problem that arises in many areas of data analysis and genomics is no exception. When the number of features or variables being analyzed exceeds the number of samples, machine learning models can make inaccurate inferences. This is often referred too as often referred to as the “curse of dimensionality” [52]. The datasets are often high dimensional in genomics data analysis. The number of

samples available maybe in order of hundreds (or even fewer), while the number of features maybe in tens of thousands. Therefore the "curse of dimensionality" proves to be a persistent challenge in genomics [52]. Thus, machine learning algorithms need to be applied to genomics data with caution. ML algorithms can be susceptible to overfitting when the number of features exceeds the number of samples. Overfitting occurs when a model is too complex and fits the training data too closely, resulting in inaccurate prediction when applied to new data. Complex models also pose the risk of selecting redundant or random features [53]. Techniques to overcome such challenges include feature selection, which identifies a subset of the most informative features for analysis, and regularization, which constrains the model to prevent overfitting, but even these methods can be data hungry to an extent [54]. Though sequencing is much cheaper than before, it is not affordable for one and all, and often, scientists are able to sequence only a few samples (tens to hundreds). This invites the need for designing novel machine learning algorithms that can overcome the challenges as well as avoid overfitting.

scRNA-seq comes with its own set of challenges [55, 56]. To reveal valuable information about the data, such as recognizing the genes that are actively expressed in specific cell types, we need to sift through the high amount of technical noise. Noise is a persistent problem that occurs with scRNA-seq [43]. This noise can arise from various sources, including low RNA input, amplification bias, and sequencing errors. The low amount of RNA that can be extracted from a single cell contributes to the high technical noise in scRNA-seq as there is a higher chance of random variation influencing the results.

Another challenge with scRNA-seq is the sparsity of the data. This refers to the fact that a high percentage of genes in a cell have zero expression [57]. This can make it difficult to identify the genes that are actively expressed in specific cell types, as the data is highly skewed towards zero.

Various computational methods for scRNA-seq data analysis have been developed to overcome these hurdles. One common approach is to use gene selection methods that focus on genes with high expression levels or high variability across cells [58,

59]. However, these methods can be vulnerable to technical noise, as genes with high variability may be more likely to be affected by noise. To overcome these challenges, newer methods are being developed that use more sophisticated statistical models to account for technical noise and sparsity in scRNA-seq data [60, 61]. These methods aim to identify more reliable markers of cell identity and function.

1.5.3 Dimensionality reduction

Feature selection in high-dimensional datasets is exceptionally challenging due to several hundred and thousands of genes profiled across multiple samples [62]. This indicates that there are a considerable number of parameters to estimate. Till date, feature selection and dimension reduction algorithms struggle with high-dimensional parameter spaces [63]. Furthermore, scRNA-seq datasets often contain technical sources of noise owing to incomplete RNA capture, polymerase chain reaction (PCR) amplification biases and/or batch effects specific to the patient or sample. Such constraints pose a big challenge for scRNA-seq analysis [43]. We must take note that although each cell contains massive gene expression profiles, only a subset of those are functionally relevant. Therefore, the efficient selection of meaningful genes from high-dimensional complex biological dataset demands novel and robust algorithms.

Dimension reduction techniques are routinely used to facilitate downstream analyses and visualization of such high dimensional datasets [64, 65]. Studying low-dimensional representation of biological datasets can often provide key insights into otherwise hidden biological mechanisms. This can, in turn, aid clinical decision making. However, non-linear dimensionality reduction techniques, though powerful are primarily used for visualization purposes only since they do not shed any light on the individual genes' identity that influences the non-linear transformation [66]. This prevents us from fully utilizing the potential of these non-linear methods for scRNA-seq analysis, as they can identify local and global structures in the data, thus exposing natural clusters and non-linear variations along the dimensions. There is significant scope in this arena to develop robust algorithms that overcome such

limitations.

1.6 A new frontier: Liquid Biopsy

Liquid biopsy is a minimally invasive diagnostic technique that involves the analysis of various biomarkers present in a patient's body fluids, such as blood, urine, and saliva. The aim of liquid biopsy is to detect and monitor diseases, such as cancer, by analyzing circulating tumour cells (CTCs), circulating tumour DNA (ctDNA), and other relevant biomolecules.

1.6.1 Need for Liquid Biopsy

Tissue biopsies are invasive since surgeries or needles may be required to obtain a sample. Repeated biopsies may also be needed in some instances. It could be that the result is inconclusive or the pathologist requires more tissue to make a diagnosis [67, 68]. There is also a higher incidence of cancer detection on repeat biopsies [69]. Though potentially helpful from a medical point of view, repeated tissue biopsies pose severe discomfort and possible infection for the patient. Thus, tissue biopsy is not a suitable method to monitor disease progression [67]. This drawback poses liquid biopsies as a potential revolution in cancer diagnostics. It is a minimally invasive method for both detecting and monitoring cancer. Liquid biopsies only require a simple blood draw or urine collection, which reduces the patient's risk of complications, pain, and infection [67]. Liquid biopsy also has a faster turnaround time than tissue biopsy, which may take weeks to process and analyze. Liquid biopsy can provide test results within days or even hours, which can help expedite diagnosis and treatment decisions.

Another advantage of liquid biopsy is that it can better capture the heterogeneity and dynamics of tumours than a tissue biopsy. Tissue biopsies may only reflect one area of one tumour at one time point. Liquid biopsies can detect multiple cancer biomarkers, such as, circulating tumour cells (CTCs), circulating tumour DNA (ctDNA), or exo-

somes [67]. ctDNA is made up of fragments of DNA that are released into the bloodstream by tumour cells. ctDNA testing examines a patient's blood to detect these DNA fragments from cancer cells [70]. CTCs are whole tumour cells that have been captured from a patient's blood sample. These cells not only reveal the presence of a tumour but also indicate that a cancer is progressing or spreading [71]. One unique advantage of ctDNA is that it provides information about the genetic makeup of the cancer cells and CTCs can provide information about their physical characteristics [72].

Due to its less invasive nature, liquid biopsy can also be used to monitor changes in disease progression and response to therapy over time by taking multiple samples at different intervals [73]. These positive features of liquid biopsy can also help encourage patients to get tested for cancer at more regular intervals and thus aid early cancer detection. Early detection plays a huge role in cancer management and recovery [74, 75]. Liquid biopsy biomarkers can also server as non-invasive alternative to predict immunotherapy response and outcomes [73]. They can be utilized to monitor treatment response and resistance as well. Thus, liquid biopsies pose to transform both cancer diagnosis and prognosis.

1.6.2 Tumour Educated Platelets (TEPs)

While ctDNA and CTCs are established liquid biopsy biomarkers, one should note that ctDNA is secreted in very low quantities by the tumour cell, and CTCs are rare and heterogeneous. These factors attribute to a higher type-II errror rate for ctDNA and CTCs. Thus, they are not reliable biomarkers for early cancer detection.

Platelets have been extensively studied for their role in blood coagulation, wound healing as well as their relationship with cancer [76]. Tumour Educated Platelets (TEPs) were developed based on observations made in the 19th century, including the fact that spontaneous coagulation is common in cancerous patients and that thrombi filled with specific tumour elements are part of tumour metastasis [77, 78]. This indicates a direct interaction between tumour cells and platelets. Further research has revealed complex interactions among megakaryocytes, platelets, and cancer, leading to the con-

cept of TEPs [79]. Platelets were first identified as detached portions of the cytoplasm of megakaryocytes in bone marrow and spleen. Circulating megakaryocytes were observed in certain pathological conditions, including cancer. Recent studies have shown that cancer causes an increase in the number of megakaryocytes in the bone marrow as a response mechanism. Platelets mediate the communication between tumour and bone via tumour-derived proteins stored in granules [76]. The relationship between platelet numbers and cancer was identified in the 1960s, with thrombocytosis being associated with a variety of diseases, including neoplasms [80]. Platelets also interact directly with circulating tumour cells. Changes in platelet numbers, proteome, size, and ratio with immune cells have been observed in the presence of cancer [81].

In the process of interacting with tumour cells, the RNA profile of the platelet gets altered, thus resulting in the platelets becoming "educated". External stimuli such as lipopolysaccharide, P-selectin ligands, and thrombin can activate kinase pathway signalling, and may also lead to "platelet education".

Platelets are fragments without a nucleus that contain pre-mRNA from megakaryocytes. Some of this pre-mRNA is spliced into mature mRNAs and used to make proteins when triggered by external signals. Therefore, platelet protein synthesis and signaling can be directly activated by tumour cells. Platelets can sequester tumour-associated biomolecules such as proteins and RNA, besides educating themselves through direct interaction with tumour cells [79].

1.7 Scope of thesis

In this dissertation, the goal has been to develop machine learning based strategies for discovering biomarkers and for diagnosis as well as personalised medicine recommendation in cancer. We have shown how leveraging machine learning tools can help us bring solutions from bench to bedside and help overcome some real world challenges. Some key aspects of clinical decision making are improving treatment efficiency, reducing adverse effects, lowering the cost for patients and care providers, diagnosing

the disease early, understanding the factors responsible for clonal subpopulation, etc. We have addressed some of these crucial components in our studies. We have also shown how machine learning algorithms can be designed intelligently to overcome real world data challenges such as - data paucity, mixed clinical data, high dimensionality, reproducibility, interpretability. Starting with the molecular intricacies revealed by *InGene* in single-cell RNA sequencing, it lays a foundation for understanding cancer at the cellular level. We then progress to address the critical need for early, accurate, and patient-friendly diagnostics. Here, we propose a non-invasive, efficient liquid biopsy diagnostic technique for NSCLC. Finally, the work culminates in the advancement of patient care through a novel staging system for Multiple Myeloma, illustrating the integration of scientific discoveries into clinical practice. Below is an anecdote of the key developments and results.

1.7.1 Finding influential genes from embeddings of non-linear dimensionality reduction techniques for single cell transcriptomes

Single-cell RNA sequencing (scRNA-seq) provides a powerful means of characterizing transcriptional heterogeneity within cells of seemingly identical phenotypes. The usual number of cells analyzed in a single-cell study ranges between a few hundreds to several thousands. Though useful for cellular phenotyping, single-cell sequencing, due to its cost, is not routinely used in clinical diagnostics. To make use of scRNA-seq in disease detection, monitoring and progression, we need targeted sequencing. Due to factors like high variability in scRNA-seq data, high dimensionality, and sparsity, traditional feature selection methods fall short in this task. Dimension reduction has emerged as a routine step in the analysis and visualization of scRNA-seq. However, non-linear dimensionality reduction techniques like t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) are primarily used for visualization purposes only since they do not shed any light on the individual genes' identity that influence the non-linear transformation. This prevents

us from fully utilizing the potential of these non-linear methods for scRNA-seq analysis, as they can identify local and global structures in the data, thus exposing natural clusters and non-linear variations along the dimensions. We developed *InGene* to overcome this limitation. *InGene* is a first-of-its-kind non-linear, unsupervised method that assigns an importance score to each expressed gene by measuring its contribution to the construction of the low-dimensional map. Gene-set enrichment analysis on top *InGene* genes for Melanoma and Human Breast Cancer shows that *InGene* genes have the best disease-gene association compared to other supervised and unsupervised methods. Thus, *InGene* can be used to obtain reliable targeted panels for scRNA-sequencing, thus reducing the cost manifold. Using a cost-effective scRNA-seq sequencing solution can prove to be a headway in personalized therapy recommendation and help make the clinical decision making process more effective. *InGene* also captures genes relevant for survival analysis and can decipher cell types accurately for different single-cell datasets. *InGene* is fast, scalable and assay-independent.

1.7.2 Building affordable liquid biopsy based on platelet transcriptome

Early diagnosis is crucial for effective medical management of cancer patients. Tissue biopsy has been widely used for cancer diagnosis, but it's invasive nature limits its application, especially when repeated biopsies are needed. Tumour Educated Platelets (TEPs) have, of late, generated considerable interest due to their ability to infer tumour existence and subtypes accurately. So far, a majority of the studies involving TEPs have offered marker-panels consisting of several hundreds of genes. Profiling large numbers of genes incurs a significant cost, impeding its diagnostic adoption. To address this, we analyzed publicly available TEP expression profiles and identified a panel of 11 platelet-genes that reliably discriminates between cancer and healthy samples.

1.7.3 Developing a risk stratification model for Multiple Myeloma patients

Over the last decade, autologous stem cell transplantation (ASCT) has emerged as the standard of care in the management of Multiple Myeloma (MM). However, the cases of early relapse (within 36 months) after the stem cell rescue remains a significant challenge. For a lot of practical purposes, it is crucial to identify whether a patient undergoing ASCT falls into the high-risk group (likely to relapse within 36 months) or a low risk one. In an effort to improve prognosis estimation and treatment efficacy for Multiple Myeloma (MM), we developed a staging system to predict the risk of relapse in MM patients undergoing Autologous Stem Cell Transplantation (ASCT). For a lot of practical purposes, it is crucial to identify whether a patient undergoing ASCT falls into the high-risk group (likely to relapse within 36 months) or a low risk one. Our designed model provides a 3-factor multivariate 2-stage staging scheme, which turns out to be highly decisive about the outcome of the stem cell rescue.

1.7.4 Synthesis and Concluding Remarks

This thesis converges the efforts from three pivotal studies, each accentuating a critical facet of oncology: molecular biology, diagnosis, and clinical management. However, beyond the confines of theoretical research and laboratory investigations, the potency of these studies lies in their translational promise. Figure 1.3 elucidates the contributions made by Chapters 2-4, and the paragraphs below elaborate the outline in greater detail.

Starting with insights into cellular genomics, we delved into the complexities of cancer at a microscopic level. *InGene* showcases the prowess of computational methodologies in oncogenomics. By extracting influential genes through embeddings from non-linear dimension reduction techniques, the study offers a data-driven approach to understanding gene regulation and influence, opening avenues for targeted molecular interventions.

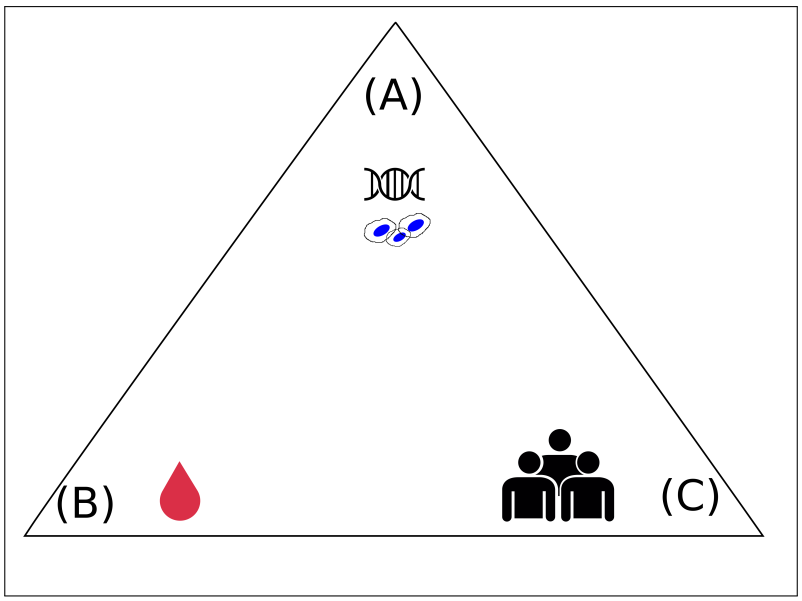

Figure 1.3: **Schematic representation of thesis.** **(A)** *InGene* helps us dissect the molecular biology of cancer by extracting influential genes through embeddings from non-linear dimension reduction techniques. **(B)** The discovery of a blood-based molecular signature encompassing 11 platelet genes improves cancer diagnosis in a non-invasive manner. **(C)** Guides clinical management of cancer patients (Multiple Myeloma) by enabling clinicians to strategize treatments based on individual relapse risks.

We then transition from dissecting molecular biology to improving diagnosis. In an era where early detection can significantly improve prognoses, the discovery of a blood-based molecular signature encompassing 11 platelet genes revolutionizes the diagnosis of NSCLC. By offering a non-invasive diagnostic method, this research has the potential to reshape clinical paradigms.

Lastly, the precision with which we manage and treat cancer patients remains paramount. The development of a staging system for Multiple Myeloma patients post-autologous stem cell transplantation is emblematic of patient-centric research. This system, grounded in clinical data, enables clinicians to strategize treatments based on individual relapse risks.

Collectively, these studies illuminate the path for future oncology research, underscoring the importance of integrating molecular insights with clinical applicability. The works presented in this thesis set the stage for the subsequent chapters, promising a journey through the multifaceted realms of cancer research.

CHAPTER 2

Finding influential genes from embeddings of non-linear dimensionality reduction techniques for single cell transcriptomes

2.1 Introduction and Motivation

Dimensionality reduction is often used to visualize high dimensional and complex data. The ability to represent multi-dimensional samples on a two-dimensional scatterplot lends a multitude of benefits. It helps us investigate the patterns of similarity between individual samples and can also aid in discovering rare entities [82]. When the dimensionality reduction technique is linear in nature, such as Principal Component Analysis (PCA), we can also decipher which features contribute to the variability of the data and thus hold higher importance. Interpreting feature importance is essential for understanding the data, model analysis and improvement. When it comes to biological research, it is exceptionally crucial that one can analyze and interpret both samples and their features (e.g. genes).

The rapid progress in next-generation sequencing (NGS) in recent years has provided many valuable insights into complex biological systems. In addition to revealing an incredible amount of information about fundamental biology and disease, high-throughput sequencing technologies are becoming increasingly versatile. They are applied to various clinical problems, ranging from cancer genomics to diverse microbial communities. NGS-based technologies for genomics, transcriptomics, and epigenomics are now increasingly focused on the characterization of individual cells. Several studies have shown that genomic alterations are more heterogeneous at the single-cell level than at the bulk level [83].

The usual number of cells analyzed in a single-cell study ranges between a few hundred to several thousand [84]. Dimension reduction techniques are routinely used to facilitate downstream analyses and visualization of such high dimensional datasets. Such methods replace several thousands of genes with a small number of latent variables while preserving important patterns in datasets. In the field of single-cell transcriptomics, Principal Component Analysis (PCA) [85], t-Distributed Stochastic Neighbor Embedding (t-SNE) [86] and Uniform Manifold Approximation Projection (UMAP) [87] are used the most [88, 89]. t-SNE and UMAP both claim to position similar cells closer together. Becht et al. [90] argued that UMAP is preferable to t-SNE because it better preserves the global structure of the data and is more consistent across runs. That being said, both tSNE and UMAP are immensely popular tools today for data visualization.

PCA uses a linear transformation to project individual cells into a latent space while maximizing the variance of the projections. On the contrary, t-SNE and UMAP identify hidden structures in data, exposing natural clusters and non-linear variations along the dimensions [86] . Following dimension reduction, an obvious question to ask is, "What are the key genes?". For PCA, often genes with the highest loadings in the first few principal components are considered to be the most influential [91, 88]. t-SNE and UMAP are non-linear dimension reduction techniques that can identify hidden data structures and expose natural clusters and non-linear variations along the dimensions. These techniques are often used to visualize high-dimensional data, but they do not have an equivalent technique to identify the most influential genes as PCA does with loadings. We propose *InGene*, a machine learning based tool for identifying genes that best explain non-linear transformations for single-cell transcriptome data.

2.2 Results

2.2.1 Overview of *InGene*

Popular feature selection techniques to detect highly variable genes are generally dispersion based methods such as - Coefficient of variation (CV) [92], Fano Factor [93] and Gini Coefficient or Gini Index [94]. They are all univariate and are calculated using the mean and standard deviation of the features. The drawback of rank ordering genes using these techniques are quite a few. Using a variance based method means that low level expression and low variability features will be under-sampled due to their low variability values. In addition, genes with higher expressions and variability values have a higher probability of showing up. Also, such methods do not capture the relative difference in expression values. For example, genes with high expressions and low variability in cells may not appear believable since their relative difference in expression value will be close to zero compared to low expression and high variability cells. Please note that dispersion-based measures of heterogeneity are often overemphasized, which can result in the selection of marker genes that are actually weakly associated with the true sub-population.

To overcome this, we believe that the cellular arrangement of the sub-population should be considered while determining essential marker genes. Thus, *InGene* is designed to utilize the structure of the sub-population during the selection of marker genes.

Dispersion based methods also do not do a good job of conveying the inherent relationships between the variance in transcript levels between cells and the variables believed to be essential for separating cell types. The methods can not fully utilize the biological knowledge of the data structure, such as co-expression or similarity structure. Therefore, methods that focus on highly variable genes based on the scores given by dispersion methods can be inadequate. To circumvent the above issues, we skipped selecting only the highly variable genes and instead chose to retain all the genes that are expressed post gene filtering in *InGene*.

Another popular technique to rank order genes is Principal Component Analysis (PCA) [95]. PCA is a dimension reduction method of transforming high dimensional data into low dimensional space. PCA provides an informative basis for attribute ranking or exploring the correlation between the genes. Though it circumvents some of the drawbacks of dispersion based methods, PCA is not very efficient when it comes to preserving local structures. Being a linear method, it also fails to perform adequately for more heterogeneous datasets.

Non-linear dimensionality reduction methods, such as t-distributed stochastic neighbor embedding (tSNE) [86] and uniform manifold approximation and projection (UMAP) [87] are used to construct an embedding of the data that preserves pairwise distances across the projected low-dimensional space. These methods are designed for the embedding of high dimensional data into some of its low dimensional projections. They can be used to obtain low dimensional representations of a set of data that preserve the relationships among high-dimensional projections are such as pairwise spatial distribution. Though much more potent in retaining the structure of the data in low dimensional space when compared to PCA and dispersion methods, non-linear dimensionality reduction methods have remained only as a mere visualization tool. Due to their non-linear nature, one is unable to decode which genes carry a heavier weightage (importance) in the construction of the lower dimensional embeddings. Thus these techniques remain underutilized when it comes to analysis. While visual interpretation does help one decipher the patterns in a dataset, it is pertinent to decode the key features influencing those patterns, especially when dealing with biological and clinical datasets.

Figure 2.1 describes how *InGene* extracts the features that best explain the cellular arrangement in the latent dimension (see Methods).

InGene was compared with popular unsupervised and supervised methods that rank orders genes for single-cell transcriptomic data - CV^2 [92], Fano Factor [93], Gini Coefficient or Gini Index [94], PCA [85], MAST [96], scGeneFit [97]. The same is evident when we compare the reconstructed tSNE/UMAP using only the top-ranked *InGene*

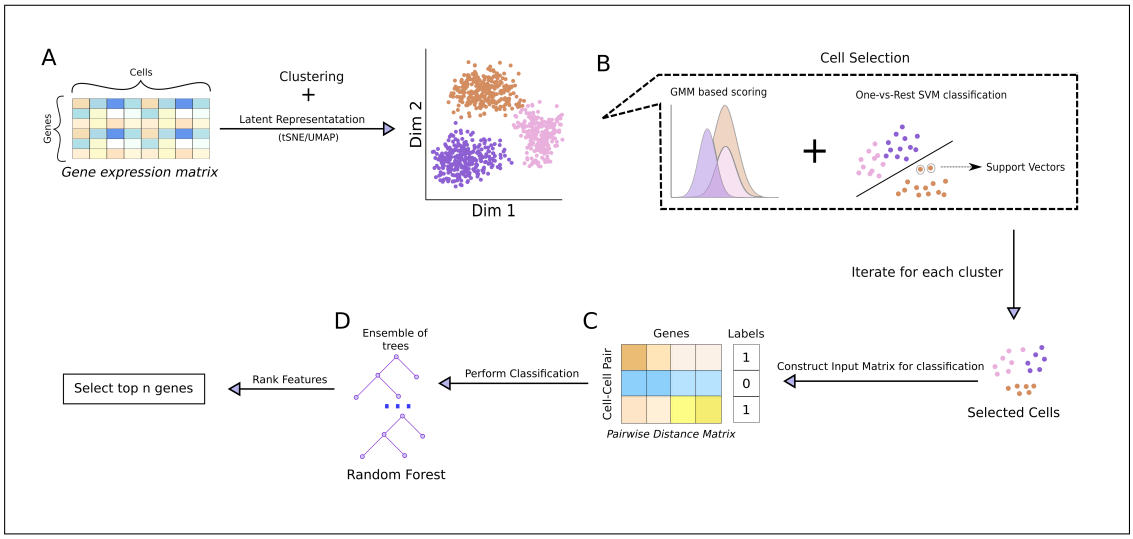

Figure 2.1: **Schematic representation of workflow.** **(A)** The gene expression data is clustered (using Seurat pipeline). The results are projected onto a 2D latent representation using tSNE/UMAP. **(B)** Cells are selected from each cluster using the latent representation. The subsampling is done using one-vs-rest SVM, combined with likelihood values from GMM. **(C)** Subsampled cells are combined pairwise to construct the feature matrix to be used for feature ranking. Each pair of cell is a new entry, and labeled as "1" if both the cells belong to same cluster or "0" otherwise. **(D)** The feature matrix is passed to Random Forest classifier, which also ranks the features (genes).

genes (max 500) along with the one that was constructed using all the genes.

The observation was consistent for all the datasets we applied *InGene* on (Figure 2.2, Figure 2.3, Figure 2.4, Figure 2.5).

2.2.2 *InGene* reveals relevant genes

InGene was evaluated on four datasets comprising various diseases, cell types, tissues, and sizes (see Methods). We accounted for some significant challenges faced in real-world data and evaluated the generalizability of the algorithm. In this work, the tested datasets span diverse applications - including but not limited to immunotherapy [98], spatial transcriptomics [99], cell type profiling [100], cell differentiation [91]. For each dataset, we demonstrate how top *InGene* genes selected via structure-preserving sampling reveal relevant genes according to the cell type, disease association, and tissue of origin.

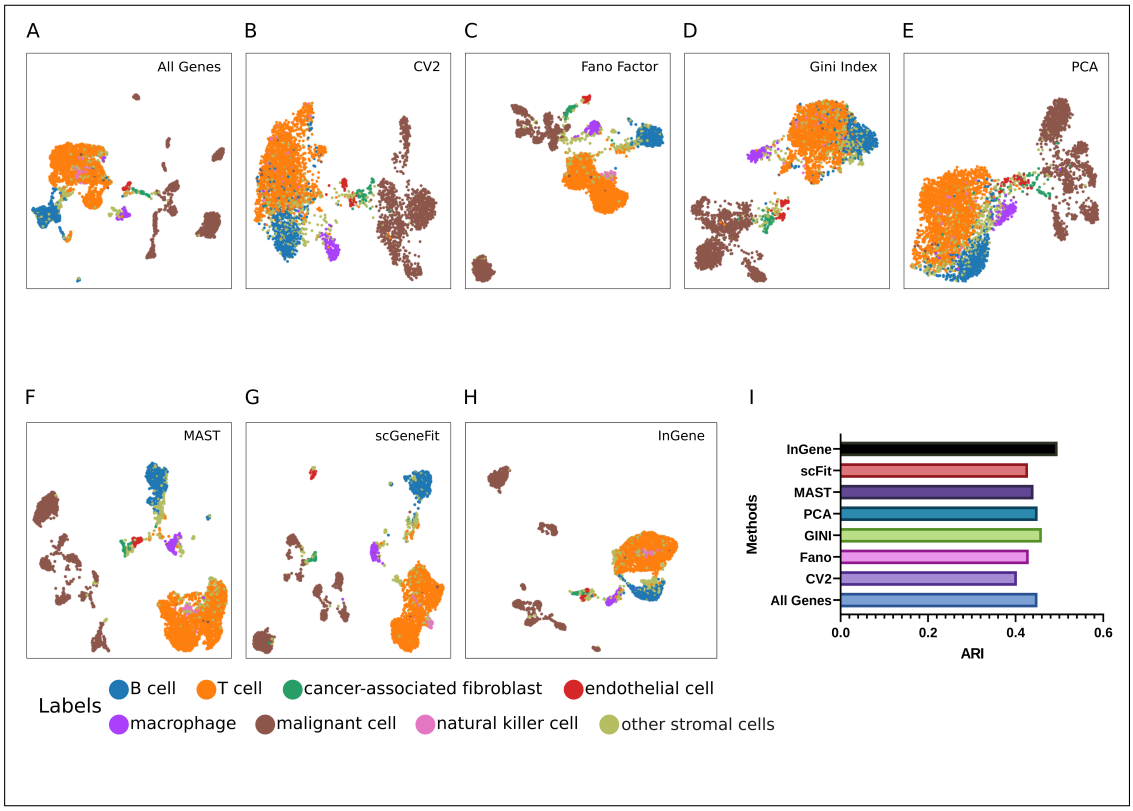

Figure 2.2: **Explaining UMAP for single-cell melanoma dataset with *InGene*.** **(A)** UMAP constructed with all the genes post-filtering. **(B)** UMAP constructed with top 500 CV2 genes. **(C)** UMAP constructed with top 500 Fano Factor genes. **(D)** UMAP constructed with top 500 Gini Index genes. **(E)** UMAP constructed with top 500 PCA genes. **(F)** UMAP constructed with top 500 MAST genes. **(G)** UMAP constructed with top 500 scGeneFit genes **(H)** UMAP constructed with top 500 *InGene* Factor genes. **(I)** ARI scores for the different methods. The gene set from each methods is used to cluster the dataset, using Leiden algorithm. The cluster labels obtained are then compared with the true labels to obtain the ARI values.

After applying the structure-preserving sampling followed by the binary classification pipeline (see Methods), we use the top 500 ranked *InGene* genes for further analysis. To maintain parity with the other feature ranking methods used for comparison purposes, we also utilize the top 500 ranked genes from those methods.

We evaluated the 500 leading genes ranked by *InGene*, CV^2, Fano Factor, Gini Index, PCA, MAST, and scGeneFit for the single-cell melanoma dataset. We found that the disease-gene association for *InGene* genes was most accurate compared to the other techniques (Figure 2.6).

The disease-gene association was captured from DisGeNET database [101] using

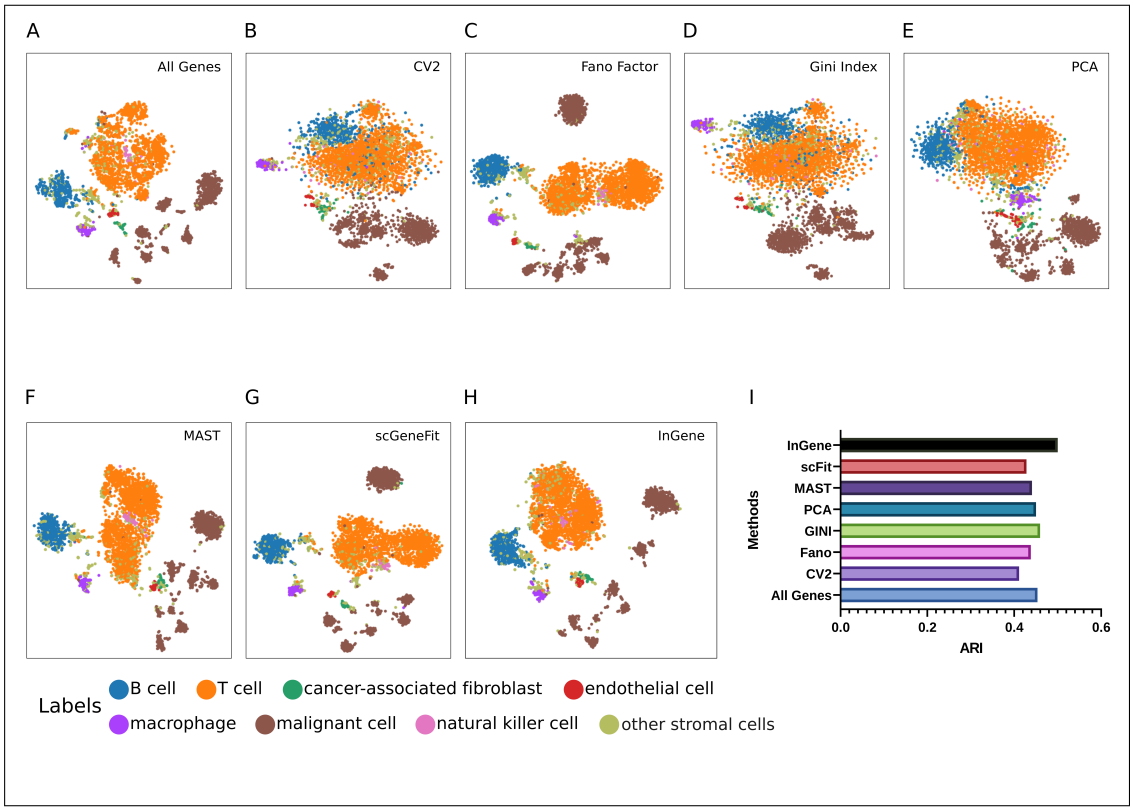

Figure 2.3: **Explaining tSNE for single-cell melanoma dataset with *InGene*.** **(A)** tSNE constructed with all the genes post-filtering. **(B)** tSNE constructed with top 500 CV2 genes. **(C)** tSNE constructed with top 500 Fano Factor genes. **(D)** tSNE constructed with top 500 Gini Index genes. **(E)** tSNE constructed with top 500 PCA genes. **(F)** tSNE constructed with top 500 MAST genes. **(G)** tSNE constructed with top 500 scGeneFit genes **(H)** tSNE constructed with top 500 *InGene* Factor genes. **(I)** ARI scores for the different methods. The gene set from each method is used to cluster the dataset, using Leiden algorithm. The cluster labels obtained are then compared with the true labels to obtain the ARI values.

EnrichR [102]. *InGene* captures key genes associated with melanoma such as - ASAH1, GNG10, HSPG2, TMSB10, XAGE1A, DSG2, etc.

- ASAH1 is necessary for melanoma tumour growth and metastasis [103, 104] It was recently established in a study [103] that ASAH1 is overexpressed in melanoma and acts as a metabolic driver. Phenotypic plasticity, the ability of cells to change their characteristics, is an essential factor in developing drug resistance and metastasis in melanoma. The plasticity involves the regulation of different transcriptional programs, including MITF, and allows melanoma cells to switch between proliferative and invasive states. The protein ASAH1, which controls the metabolism of sphingolipids, was found to play a role in this switch in melanoma cells [103]. Experiments showed that the ASAH1 could be used as a target for melanoma therapy, as it increases the effectiveness of BRAF inhibitors [103].

- GNG10, present in the family of G-proteins, is mutated in melanoma [105]. Ac-

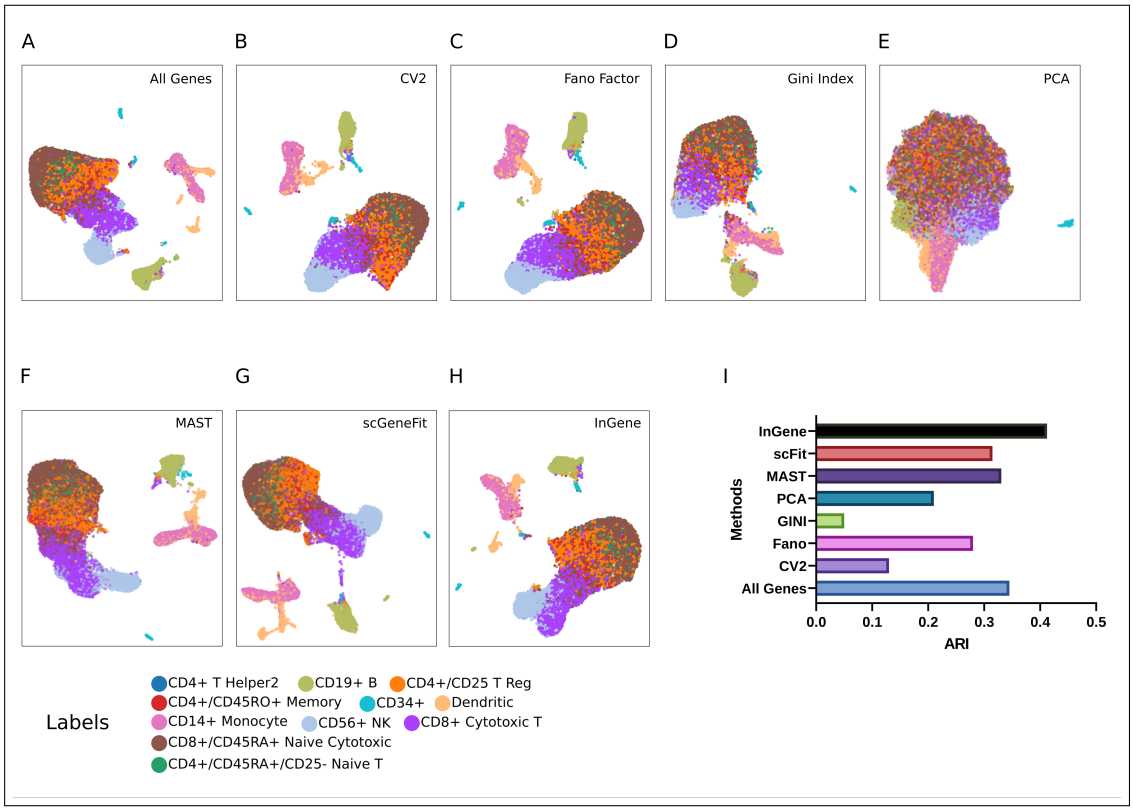

Figure 2.4: **Explaining UMAP for single-cell PBMC 68K dataset with *InGene*. (A)** UMAP constructed with all the genes post-filtering. **(B)** UMAP constructed with top 500 CV2 genes. **(C)** UMAP constructed with top 500 Fano Factor genes. **(D)** UMAP constructed with top 500 Gini Index genes. **(E)** UMAP constructed with top 500 PCA genes. **(F)** UMAP constructed with top 500 MAST genes. **(G)** UMAP constructed with top 500 scGeneFit genes **(H)** UMAP constructed with top 500 *InGene* Factor genes. **(I)** ARI scores for the different methods. The gene set from each methods is used to cluster the dataset, using Leiden algorithm. The cluster labels obtained are then compared with the true labels to obtain the ARI values.

cording to a study published in 2010, GNG10 was found to have a high rate of non-synonymous mutations (changes to the DNA sequence that alter the amino acid sequence of the protein) in melanoma tumours. Moreover, these mutations were associated with the progression of the disease [105]. The finding suggests that GNG10 may be involved in the development and progression of melanoma and that these mutations may be useful as markers for predicting the severity of the disease.

- HSPG2 is frequently mutated in melanoma and is over-expressed [106]. Immune checkpoint inhibitor (ICIs) therapy improves the survival outcome of advanced melanoma patients. Zhang et. al [107] showed an association between melanoma patients with HSPG2 mutations had an improved ICI outcome. The authors found melanoma patients with mutations in the HSPG2 gene had better outcomes when treated with immunotherapy (ICI) compared to other patients. They also found that these patients had higher levels of immune cells that respond to cancer (response immunocytes), lower levels of immune cells that suppress the immune re-

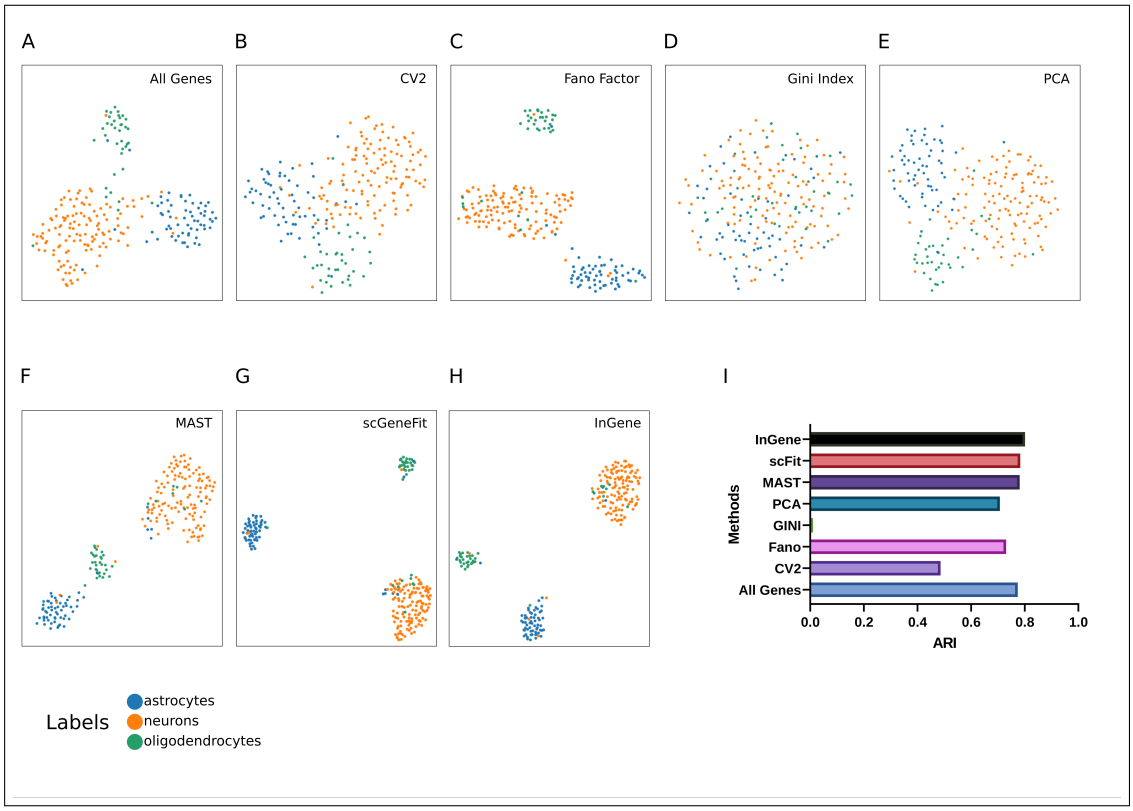

Figure 2.5: **Explaining UMAP for single-cell Darmanis dataset with *InGene*.** **(A)** UMAP constructed with all the genes post-filtering. **(B)** UMAP constructed with top 500 CV2 genes. **(C)** UMAP constructed with top 500 Fano Factor genes. **(D)** UMAP constructed with top 500 Gini Index genes. **(E)** UMAP constructed with top 500 PCA genes. **(F)** UMAP constructed with top 500 MAST genes. **(G)** UMAP constructed with top 500 scGeneFit genes **(H)** UMAP constructed with top 500 *InGene* Factor genes. **(I)** ARI scores for the different methods. The gene set from each methods is used to cluster the dataset, using Leiden algorithm. The cluster labels obtained are then compared with the true labels to obtain the ARI values.

sponse (suppression immunocytes), more mutations in their cancer cells, and increased activity in pathways related to the immune system's response to infection (interferon response-relevant signaling pathways). Based on these findings, the authors concluded that HSPG2 mutations may be a predictor of a good response to immunotherapy and may be used to guide treatment decisions for melanoma patients.

- TMSB10 is a well-established gene signature for predicting survival of metastatic melanoma patients [108, 109]. The authors of a study [110] found that the protein TMSB10 was associated with the ability of melanoma cells to spread to other parts of the body (metastasis) in animal models and in fresh samples of human melanoma tissue. They also found that TMSB10 was not consistently related to the spread of cancer or the severity of disease in other types of cancer. Based on these findings, the authors concluded that TMSB10 might be a helpful marker for predicting the severity of cutaneous melanoma.

- DSG2 promotes vasculogenic mimicry (VM) in melanoma [111]. Overexpres-

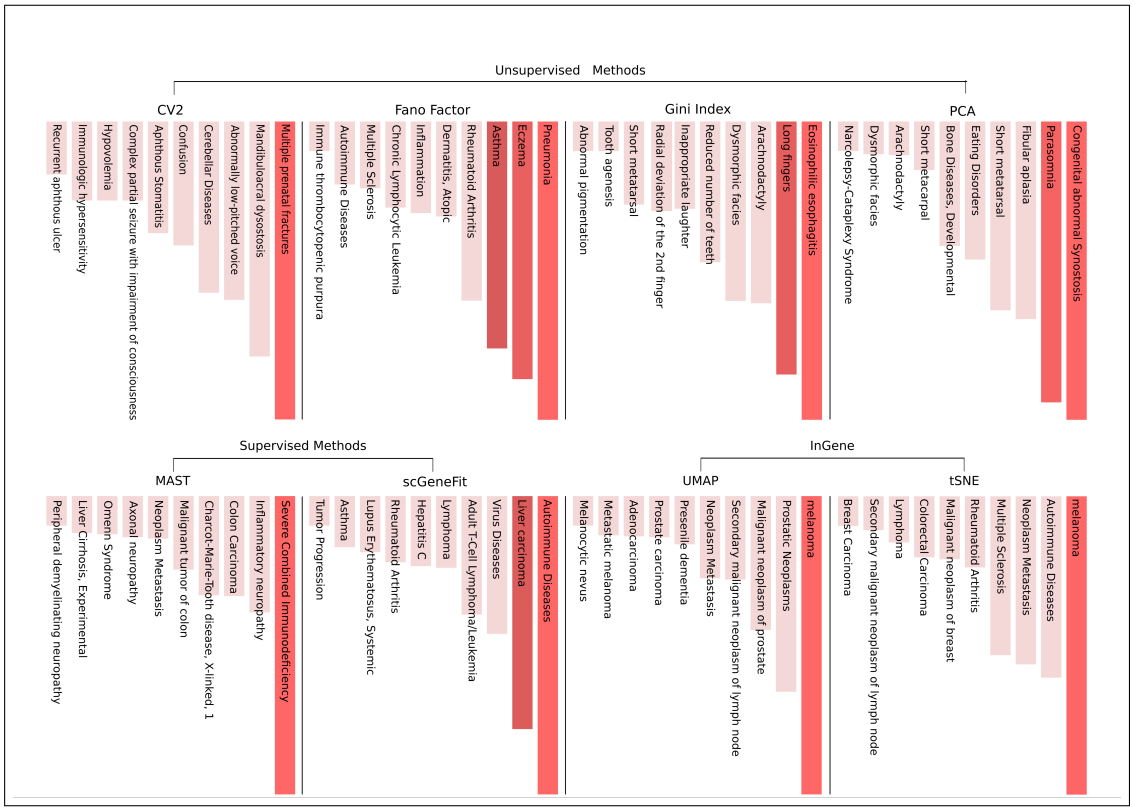

Figure 2.6: **Disease-Gene association for top 500 genes from each method for GSE72056**. *InGene* genes obtained from 2D UMAP and tSNE embeddings correctly rank Melanoma as the associated disease, with the highest significance. Other methods - both supervised and unsupervised fail to capture the disease-gene association for the dataset.

sion of DSG2 is associated with poor clinical outcomes in melanoma patients [112, 111]. The study [111] establishes that DSG2 is expressed in primary and metastatic melanoma tissue but not in normal melanocytes. DSG2 plays a critical role in regulating the angiogenic activity of Endothelial Cells (ECs) and circulating endothelial progenitor cells (EPCs). The authors [111] have shown that DSG2 plays a similar cell-intrinsic role in melanoma by regulating VM and thus may prove to be a new treatment approach for melanoma.

2.2.3 *InGene* captures spatially differential genes

The spatial organization and heterogeneity of gene expression within a tissue have significant biological effects on the properties of the tissue. Regular transcriptome analyses using bulk or single-cell sequencing do not capture high-resolution spatial heterogeneity. Using these techniques, the rich spatial information about gene expression is lost. Recent developments in spatial transcriptomics capture spatial information by using

DNA barcodes to distinguish different spots in the tissue. Detecting spatially differentially expressed genes can help reveal new marker genes, pathways, and molecular mechanisms and could also prove to be therapeutic targets. We used the spatial expression profile of the publicly available spatial gene expression for Human Breast Cancer (Block A Section 1) from the 10X genomics support website [99]. Notably, *InGene* has the highest overlap with differentially expressed genes from SpatialDE [113], a well-known method specially designed to reveal differentially expressed genes in spatial transcriptomics data (Figure 2.7[E]).

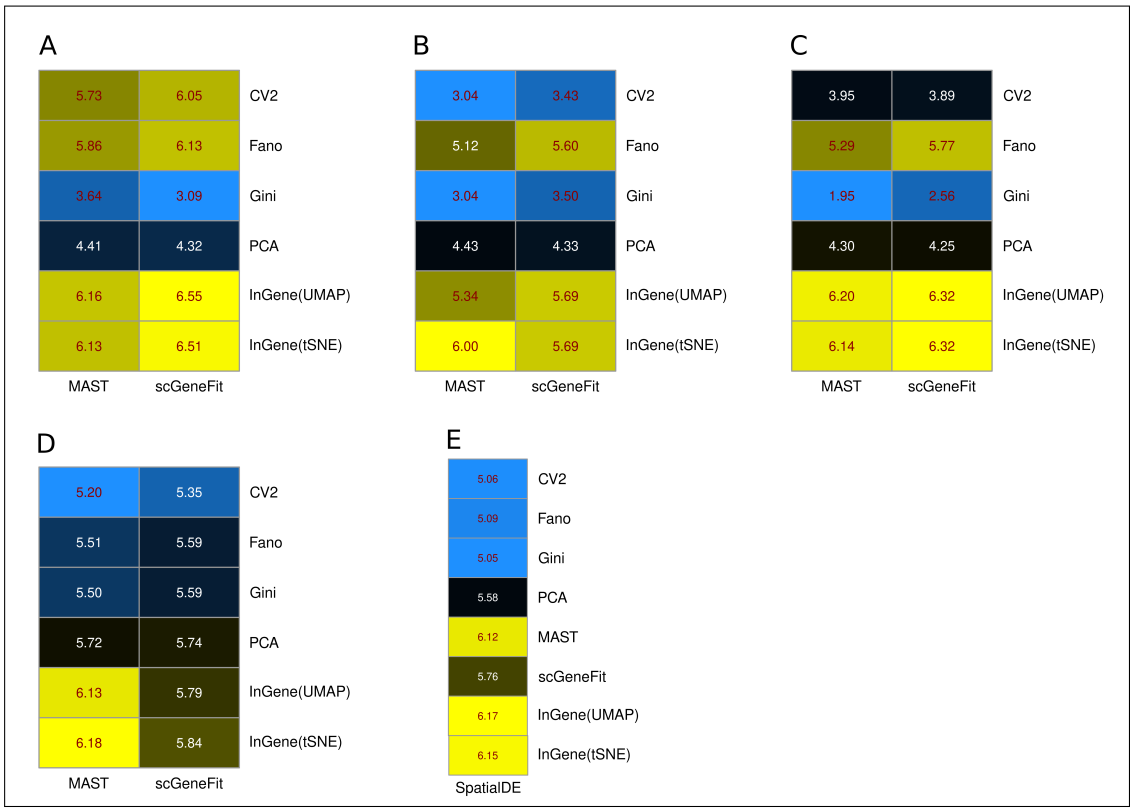

Figure 2.7: ***InGene* performs closest to supervised methods. Heatmap (log2-scale) shows that the leading 500 *Ingene* genes has the highest intersection with the supervised methods - MAST, scGeneFit. The comparison is performed for all four datasets**. **(A)** Overlap of leading 500 genes from each unsupervised method with those from MAST, scGeneFit, for PBMC 68K dataset. **(B)** Overlap of leading 500 genes from each unsupervised method with those from MAST, scGeneFit, for Melanoma dataset. **(C)** Overlap of leading 500 genes from each unsupervised method with those from MAST, scGeneFit, for Darmanis dataset. **(D)** Overlap of leading 500 genes from each unsupervised method with those from MAST, scGeneFit, for Human Breast Cancer dataset. **(E)** Overlap of leading 500 genes from each method with those from SpatialDE, for Human Breast Cancer dataset.

InGene successfully captured relevant differentially expressed genes in an unsuper-

vised manner without any input of spatial information. Some relevant differential genes captured by *InGene* are - MYC, EIF4A2, HSPG2, JPT1, NSMCE2.

- MYC is a protein that plays a role in controlling the growth and division of cells. When MYC is not regulated correctly, it can contribute to the development of various types of cancer, including breast cancer. MYC has been found to be active in the progression of breast cancer, and when combined with the loss of another protein called BRCA1, it can lead to the development of a specific type of breast cancer called basal-like breast cancer. MYC may also contribute to resistance to treatment for cancer, as it is influenced by the estrogen receptor and epidermal growth factor receptor pathways [114, 115].

- EIF4A2 plays a role in mRNA translation and has been implicated in cancer development and progression of breast cancer [116]. According to a study by Liu et al., eIF4A2 mRNA levels were significantly higher in breast cancer tissues resistant to paclitaxel treatment compared to those that were sensitive to paclitaxel [117]. Additionally, experiments have shown that reducing the amount of eIF4A2 protein in triple-negative breast cancer cells led to a decrease in cell proliferation and an increase in cell death (apoptosis) [117]. These findings suggest that eIF4A2 may be a potential target for breast cancer treatment, particularly for triple-negative breast cancer.

- HSPG2, also known as perlecan, is a protein involved in the extracellular matrix and overexpressed in some types of cancer, including breast cancer [118, 119]. Several studies have demonstrated that HSPG2 overexpression was associated with invasion, metastasis, and an inferior survival outcome in triple-negative breast cancer [120]. HSPG2 has been reported as a novel target for breast cancer [121, 122].

- JPT1, also known as HN1 was found to be upregulated in breast cancer tissues [123]. The authors of the study concluded that HN1 may contribute to the progression of breast cancer by increasing the expression of the protein MYC and that it may be a potential target for the treatment of breast cancer. Another study published in 2018 found that the HN1-like gene (HN1L) is frequently altered in breast cancer, particularly in triple-negative breast cancer (TNBC), and is associated with a poorer prognosis for patients [124]. This study found that reducing the amount of HN1L protein in breast cancer cells led to a decrease in the population of breast cancer stem cells, made chemotherapy more effective, and slowed the progression of TNBC in animal models. Additionally, the researchers found that certain gene patterns were linked to shorter disease-free survival for TNBC patients. These findings suggest that HN1L may be a potential target for the treatment of TNBC.

- NSMCE2 was identified as a novel super-enhancer-associated gene – NSMCE2 in a recent study [125]. The authors found that high levels of NSMCE2 are associated with a poor prognosis for breast cancer patients. They also found that disrupting certain regulatory regions called super-enhancers can increase the production of NSMCE2 and that high levels of NSMCE2 are linked to a poor response to chemotherapy, particularly in patients with aggressive triple-negative and HER2-positive breast cancer. Furthermore, decreasing the amount of NSMCE2 in breast cancer cells made them more sensitive to chemotherapy treatment.

2.2.4 *InGene* as an alternative unsupervised method to obtain differential genes

The above results show that *InGene* efficiently reveals relevant genes by learning the factors that dominate the latent representation of the dataset. Therefore, *InGene* can be used as an alternative to popular unsupervised methods to obtain differentially expressed genes (DEGs), such as t-test and wilcoxon test. It should be noted that the popular methods for DEG selection for single-cell RNA-seq data are supervised in nature (MAST [96], scGeneFit [97]). Notably, existing DE gene selection methods have been restricted to two-group comparisons. This is a hindrance, especially when it comes to single-cell data, which often has multiple groups in the dataset. Comparing only two groups at a time leads to post-processing overhead. It also does not allow the user to look at the comprehensive overall result because taking a union of all DE genes from each two-group comparison is not an effective solution to this end. *InGene* overcomes the shortcoming by ranking DE genes representative of each group present in the dataset. We achieve so by converting the multi-class classification problem to a binary classification problem (see Methods). Thus, the user does not need to perform additional steps for a multi-group comparison when using *InGene*. We also demonstrate in our experiments that the top 500 *InGene* genes lead to a higher classification accuracy (adjusted rand index (ARI)). *InGene* works efficiently on both small and large datasets. We observed that *InGene* genes efficiently describe the associated cell types for both PBMC68K (68000 cells) and human brain data (466 cells) [100, 91].

2.3 Methods

2.3.1 Datasets

To evaluate *InGene*, we applied the method to four publicly available single-cell datasets spanning different sizes, cell types, and diseases. The following section provides a brief description.

We used single-cell RNA sequencing (scRNA-seq) of Melanoma patient tumours [98] comprising 4,645 single cells, profiling malignant, immune, stromal and endothelial cells. Post filtering, 11217 genes and 4513 were retained. We also tested if *InGene* is capable of capturing spatially differential genes. To this regard, spatial gene expression of breast cancer specimens profiled via STseq using the Visium platform of 10x Genomics (Pleasanton, CA, USA) [99] consisting of 3813 spots, comprising basal, stroma, macrophage, luminal, mesenchymal, endothelial. B-Cell and T-cells were used. Post filtering 3209 genes and all the spots were retained.

We used two additional datasets to determine the scalability of *InGene*. The first is scRNA-seq data consisting of approximately 68000 fresh peripheral blood mononuclear cells (PBMCs) collected from a healthy donor [100]. Single-cell expression profiles of 11 purified sub-populations of PBMCs are used as a reference for cell type annotation. Next, we wanted to determine if the algorithm is effective on datasets with a few cells. Darmanis et al. used single-cell RNA sequencing on 466 cells to capture human brain transcriptome diversity at the single-cell level [91]. We filtered the three most prevalent cell types, namely neurons, astrocytes and oligodendrocytes, for the assessment. We were left with 227 cells and 14692 genes post-filtering.

2.3.2 Preprocessing of scRNA-seq data

Given a raw count data matrix, poorly expressed genes were first filtered out. Cells with a low number of detected genes were also ignored. A cell needs to have at least 200 genes expressed to be retained. Similarly, genes that were not expressed in more than one percent of the cells were filtered out. The thresholds may be modified depending on the size and dimension of the dataset. Seurat was used for data normalization and scaling. Of note, we did not limit scaling the data with respect to only the top variable genes (the default is top 2000 CV^2 genes in Seurat). Instead, we chose to retain all the genes post-filtering.

2.3.3 Gaussian mixture modeling of single cells and structure preserving sampling

Once the count matrix is filtered, normalized and scaled, we utilize the Seurat pipeline to cluster the dataset using the Leiden algorithm [126]. The clustered data is then projected on the latent dimensions (tSNE/UMAP). Once we obtain the two-dimensional map, we hypothesize that each single cell x is populated from a mixture of k bivariate Gaussians,

$$f(x \mid \lambda) = \sum_{i=1}^{k} \tau_i \mathcal{N}\left(x \mid \mu_i, \Sigma_i\right) \tag{2.1}$$

where $\tau_i, i = 1, \ldots, k$, are mixture weights that add up to 1, $\mathcal{N}\left(x \mid \mu_i, \Sigma_i\right), i = 1, \ldots, k$, are the components of bivariate Gaussian densities, each of which is characterized by a mean vector μ_i and covariance matrix Σ_i.

These parameters are collectively represented by λ, where $\lambda = \{\tau_i, \mu_i, \Sigma_i\}$. The value of k is equivalent to the number of clusters determined by the Leiden algorithm. The values of μ_i and Σ_i are calculated in a supervised manner for each cluster i using the MclustDA function from the mclust package [127]. Thus, with a given μ_i and Σ_i for a cluster i, we assign each cell of the cluster i a probability or likelihood of belonging to it. We then subsample cells from each cluster by leveraging the likelihood values. If a cluster i contains n cells, we choose $min(p, n)$ cells with the highest likelihood as well as $min(p, n)$ cells with the lowest likelihood of belonging to cluster i, where the value of p typically ranges between 10 to 20. In parallel, we perform Support Vector Machine (SVM) [128] classification using the one-against-all (one-vs-rest) strategy. Given k clusters, k binary classifiers are generated. Each classifier is responsible for distinguishing a cluster i from the remaining clusters. Let us consider that cells belonging to the current cluster i are given the label $+1$, and cells belonging to all the clusters are given the label -1. SVM looks for a hyperplane (Equation 2.2), which separates the data from classes $y_\mathbf{i}$ ($+1$ and -1) with a maximal margin.

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{2.2}$$

where $\mathbf{w}$ is the normal vector to the hyperplane and b is an offset. The margin maximization solves a convex quadratic optimization problem, in which the resulting decision boundary or hyperplane is given by Equation 2.3:

$$f(\mathbf{x}) = \sum_{\mathbf{x_i} \in \mathbf{S}V} y_\mathbf{i} \alpha_i \mathbf{x_i} \cdot \mathbf{x} + b \tag{2.3}$$

where the constants α_i are called Lagrange multipliers and are determined in the optimization process. Here SV corresponds to the set of support vectors, samples for which the associated Lagrange multipliers are larger than zero. These samples are those closest to the optimal hyperplane are the ones we choose to focus on. We select $min(p/2, s)$ SVs from cluster i with the highest likelihood, where s is the total number of support vectors belonging to cluster i. In this manner, we select a few representative cells for each cluster.

Such a selection approach enables us to implement a structure-preserving subsampling. The method also allows for the efficient integration of cells across all clusters. The idea is that the cells that are the support vectors will represent the boundary or marginal cells, and the cells with the highest likelihood shall represent centrally located cells. Thus, features learnt from these cells together should be representative of the whole cluster. Since we are performing subsampling with only a few cells per cluster (approximately 20 to 30), the method is size agnostic and can scale well. Once we have all the selected cells, we construct a binary classification problem to score each gene.

2.3.4 Construction of a supervised binary classification problem

A new strategy was used to identify the key genes from a low-dimensional projection of scRNA-seq data. Instead of contrasting each group of cells to the rest, we modelled the differences between pairs of nearby cells to pairs of cells that were far off, regardless of their clusters of origin. A supervised bi-class classification problem was constructed for the same.

From each cluster, i, we utilize the subsampled cells from it. Input space for the classification problem consisted of a total of $2M$ data points, where each data point $d_i \in \mathbb{R}^{|G|}$ was created from a pair of cells (m, n). G here denotes the set of all expressed genes. Values of d_i are determined as follows.

$$d_i = |e_g^m - e_g^n| / |e_g^m + e_g^n| : g = 1, \ldots, G \tag{2.4}$$

where e_g^x and e_g^y are expression of a gene $g \in G$ in cells m and n respectively. Each data point d_i was assigned a label $y_i \in \{N, F\}$, following a simple scheme. If both cells in a pair associated with a certain data point d_i shared the same cluster identity, y_i was set N, whereas F otherwise. We constructed M data points of each type to balance the data.

The above-described transformation is a unique way to link high dimensional space spanned by several thousands of genes to a corresponding latent space. We hypothesized that genes that best explain the proximity among pairs of cells in a latent space are the ones that influence the process of dimension reduction the most.

2.3.5 Gene selection using Random Forest

For classification, we chose the RandomForests (RF) algorithm [129] for the following reasons: 1. RF provides a Gini importance score for each variable (a gene in this case) by observing its contribution to constructing several decision trees on bootstrap training sets. 2. Any parametric assumptions do not substantiate it. 3. RF is a non-linear classifier. 4. The transformation depicted by Equation 2.4 results in folded distributions which may turn out to be incompatible with regression-based methods. For example, suppose the log-transformed expression of a gene follows a normal distribution. In that case, the absolute difference in its expression between a pair of cells will follow a folded normal distribution. Regression-based methods, in this scenario, are not guaranteed to produce stable results. To build the RF models, we employed a fast and memory-efficient implementation of the algorithm, distributed through an R package

called ranger [130]. Out Of Bag Error (OOBE) reported by ranger was considered as a measure of the goodness of a model. In our cases, OOBE varied between 2-5%. Finally, all expressed genes were prioritized based on the Gini importance scores. Important to note that RF models were never used for class prediction. Our main objective was to obtain the importance scores for a supplied set of genes, which was attained with the completion of the training phase.

2.3.6 Evaluation of genes

Suppose top genes suggested by *InGene* are principal determinants of the low-dimensional map of single-cell transcriptomes. In that case, we hypothesize that relative Euclidean distances between data points spanned by these genes would be preserved in the low-dimensional space. Spearman's rank correlation coefficient between Euclidean distances was computed across cell pairs. Given a set of influential candidate genes, this simple measure comprehends their ability to reconstruct the distances observed in a latent space. For ease of interpretation, we termed this measure Reconstruction Accuracy or RA. We compared four additional unsupervised gene ranking methods (namely - CV^2, PCA, Fano Factor, Gini Index) and two supervised (MAST and scGeneFit [96, 97]) gene ranking methods with *InGene* in this regard. We extract the top 1000 genes as per the rankings determined by each of these methods and compare their RA against the top 1000 *InGene* genes. RA scores were also computed on randomly sampled gene sets of length varying from 10 to 1000 over 20 iterations. This gives us a measurement of how the methods perform against random selection.

2.3.7 Assessing the relevance of selected genes

We selected 500 leading genes from each of the following methods: CV^2, Fano Factor, Gini Index, PCA, MAST, scGeneFit, and *InGene*. To evaluate the biological significance of these gene sets, we utilized techniques such as disease gene association, cell-type identification, and survival analysis. To identify the disease gene association

and cell types, we utilized the EnrichR web server. EnrichR is a widely used tool that allows users to input a list of genes and then extracts information from various well-known databases to provide insights on the collective functions of the gene lists. For survival analysis, we employed another reputable web server, GEPIA 2. GEPIA 2 can perform survival analysis on multiple genes using both tumour and normal samples from the TCGA and GTEx databases. To examine if the 500 top-ranked genes from the GSE72056 (single-cell melanoma dataset) are indicative of survival analysis, we selected the SKCM database in GEPIA 2. We used the gene-set from each of the unsupervised methods CV^2, Fano Factor, Gini Index, PCA, and *InGene* in GEPIA 2. The output from GEPIA 2 includes Kaplan Meier (KM) survival curves for each set of input sets. Also, it provides the combined significance (p value) of the gene sets in the survival of cancer patients for the chosen cancer type. For both EnrichR and GEPIA 2, we used the official gene symbols as input. Any gene symbols that did not align with the pre-existing symbols in the internal databases of EnrichR and GEPIA 2 were eliminated from the analysis.

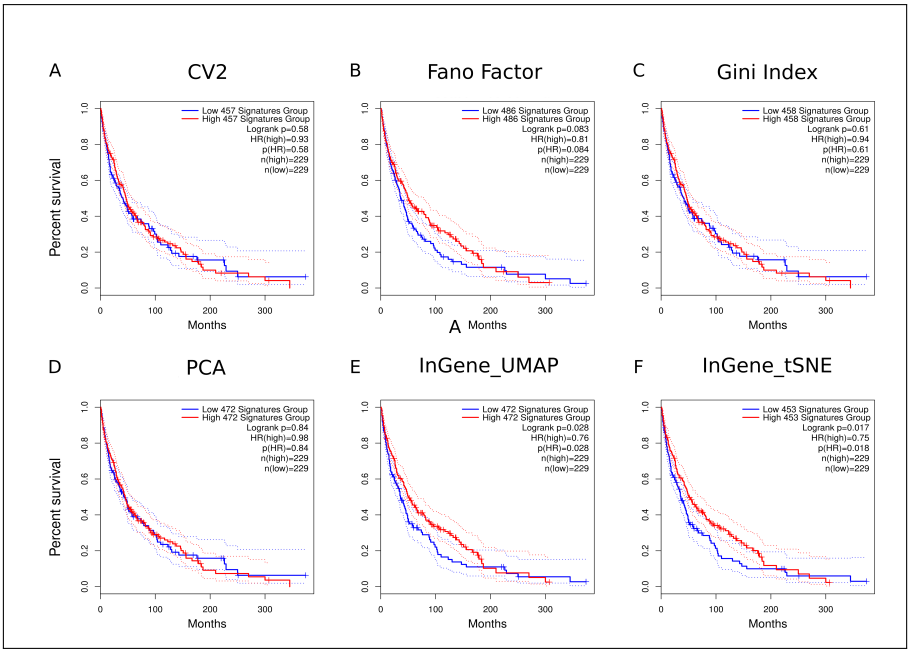

Figure 2.8: **Kaplan-Meier (KM) plots for TCGA-SKCM dataset with top 500 genes obtained from each method**. **(A)** Survival analysis (KM) on TCGA-SKCM dataset with top 500 CV2 genes obtained from melanoma dataset (GSE72056) p-value is 0.58. **(B)** Survival analysis (KM) on TCGA-SKCM dataset with top 500 Fano-Factor genes obtained from melanoma dataset (GSE72056) p-value is 0.08. **(C)** Survival analysis (KM) on TCGA-SKCM dataset with top 500 Gini genes obtained from melanoma dataset (GSE72056) p-value is 0.61. **(D)** Survival analysis (KM) on TCGA-SKCM dataset with top 500 PCA genes obtained from melanoma dataset (GSE72056) p-value is 0.84. **(E)** Survival analysis (KM) on TCGA-SKCM dataset with top 500 *InGene* genes obtained using 2D UMAP embeddings of melanoma dataset (GSE72056) p-value is 0.02. **(F)** Survival analysis (KM) on TCGA-SKCM dataset with top 500 *InGene* genes obtained using 2D tSNE embeddings of melanoma dataset (GSE72056) p-value is 0.017.

Figure 2.8 shows that *InGene* genes are predictive of survival for melanoma patients.

Noteworthy, we also assessed the speed and scalability of *InGene*. As the number of cells increase, the time taken by *InGene* increases linearly 2.9.

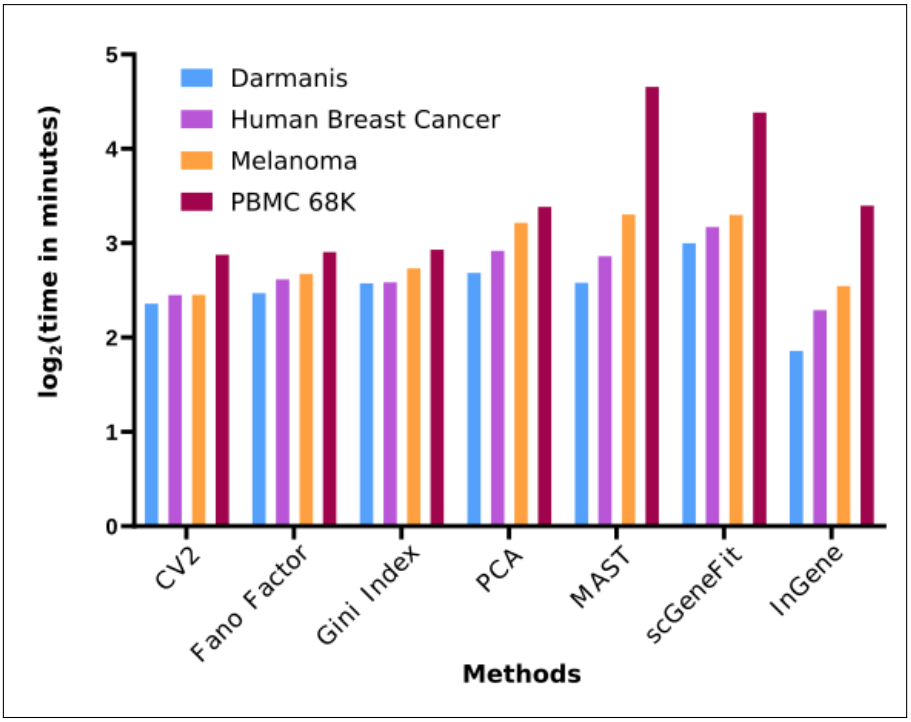

Figure 2.9: ***InGene* is scalable**. Run time recorded for each method (CV2, Fano Facor, Gini Index, PCA, MAST, scGeneFit, *InGene*) while varying the number of cells from ∼200 to ∼68K.

2.4 Discussion and Conclusion

InGene offers a robust method for analyzing scRNA-seq data without the need for pre-labeled training data and innovatively assigns importance scores to genes, enhancing understanding of cellular heterogeneity and disease-related gene identification. While these are crucial areas to address, the limitations of *InGene* must also be discussed. Computational methods like *InGene* may face challenges in accurately interpreting data due to the inherent biological variability among samples. This variability can arise from differences in tissue types, disease states, or even individual genetic differences. Since *InGene*'s accuracy is contingent upon the effectiveness of UMAP and tSNE, any limitations or inaccuracies inherent in these methods directly impact *InGene*'s performance. Considering that no dimensionality reduction technique is flawless, especially in the context of complex and often noisy biological data, this reliance can potentially be a limitation. Given these complexities, extensive validation across a wide range of datasets is crucial to ensure the robustness and generalizability of *InGene*. *InGene*, in

principle, can be applied to any linear or nonlinear dimension reduction method to extract relevant genes. *InGene* can also be used as an alternative to differential expression analysis since :

- The associated gene ranking method is free of any parametric assumption, which makes *InGene* assay-independent.
- *InGene* does not require each group of cells to be compared against the rest. Instead, it poses the whole problem of cell type-specific gene finding as a single bi-class classification problem.

InGene tackles the variability and complexity of gene expression in cancer, offering a more nuanced understanding at a cellular level. We now progress to a patient-centric approach in the next chapter. We pave the way for developing novel diagnostic methods crucial for early and accurate cancer detection, thus encompassing a broader spectrum of cancer patient care.

Code availability

InGene software package is available at: https://github.com/cgiiitd/InGene.

Data Availability

The study uses various publicly available scRNA-seq datasets. The PBMC data that support the findings of this study is available from `https://support.10xgenomics.com/single-cell-gene-expression/datasets`. The melanoma dataset can be accessed at the GEO under accession code GSE72056. The gene expression of breast cancer specimens profiled via STseq using the Visium platform of 10x Genomics is available from `https://support.10xgenomics.com/spatial-gene-expression/datasets/1.0.0/V1_Breast_Cancer_Block_A_Section_1`. The single-cell human brain data can be obtained from GEO under the accession code GSE67835.

Declarations

There are no conflicts of interest.

CHAPTER 3

Developing machine learning based strategies to design a non-invasive, blood-based inexpensive, cancer screening technology

3.1 Introduction and Motivation

Cancer is the second leading cause of death worldwide, accounting for an estimated 9.6 million deaths globally in the year 2018. Of these, 2.09 million cases were reported for lung cancer. Current omics technologies are revolutionizing the field of precision oncology by allowing the tailoring of treatment according to patient's tumour molecular profiles. Presently, the gold standard for diagnosis of cancer usually involves tissue biopsy-based molecular profiling of tumours. Although tissue-based genetic profiling of tumours has represented a probable and promising tool in cancer classification and management, in recent times, it has become more comprehensible that a single solid tissue biopsy cannot give a spatiotemporal snapshot of a tumour. It fails to capture the genetic heterogeneity of a disease, thus hampering the accuracy of the test. Furthermore, apart from being invasive in nature, these tests pose limitations of repeated sampling, sensitivity, accuracy and patient risk. Therefore, these tests are generally suboptimal and suffer setbacks. Early diagnosis is crucial for effective medical management of cancer patients. The most used method for cancer detection is tissue biopsy. Though widely used, its invasive nature limits its application, especially when repeated biopsies are needed. Over the past few years, explorations have led to the discovery of various blood-based biomarkers, such as ctDNA, cfDNA, and Tumour Educated Platelets (TEPs). TEPs, a relatively new discovery, are a promising and viable blood-based biomarker due to reduced type-II error and their ability to accurately infer tumour existence and subtype. So far, a majority of the studies involving TEPs have offered marker panels consisting

of several hundreds of genes. Profiling large numbers of genes incurs a significant cost, impeding its diagnostic adoption. As such, it is important to construct minimalistic molecular signatures comprising a small number of genes. NSCLC, the most prevalent form of lung cancer, is largely asymptomatic in its early stage. The majority of its detection takes place at an advanced stage when the disease has spread widely to distant organs. As such, the development of affordable early diagnostic tests plays a major role in improved management of the disease. For NSCLC, some studies have shown up to 100% false positives CTC detection rates in patient samples [131]. Jenkins and colleagues reported false-negative rates up to 50% in patients with intra-thoracic limited (M1a) disease while using a ctDNA-based method [132]. A recent study by Best et al. [79] revealed significant changes in platelet transcriptomes between cancer patients and healthy individuals, leading to the new tumour-tour-educated platelets (TEPs) concept. The dramatic changes in platelet transcriptome have since been linked to the cross-talk between tumour cells and platelets [133]. Using $\sim$1000 variable genes, the authors reported 96% accuracy in distinguishing localized and metastatic tumours of six major cancer types from healthy cases [79]. A study by Best and colleagues [79] showed that TEPs are substantially more accurate in predicting the existence of cancer, with false-negative and false-positive rates recorded as 4% and 8%, respectively. In an independent study focusing on Non-Small Cell Lung Cancer (NSCLC), the authors designed a classification model derived from $\sim$ 1600 genes. They reported an overall accuracy of 88% for late-stage cancer and 81% for locally advanced cancer by employing statistical and machine learning-based techniques [134]. More recently, Sheng and colleagues leveraged the RNA sequencing (RNA-seq) dataset published by Best et al. [134] to achieve 88.9% accuracy for NSCLC classification with a mere 48 genes. Their work highlighted the scope of retaining predictability with a concise gene panel, thereby inspiring its potential diagnostic use [135]. While such informative explorations extend the field, they seldom see materialization due to a lack of validation.

To address the aforesaid challenges, we analyzed publicly available TEP expression profiles and identified a panel of 11 platelet genes that reliably discriminates between cancer and healthy samples. To validate its efficacy, we chose non-small cell lung

cancer (NSCLC), the most prevalent type of lung malignancy. NSCLC often remains asymptomatic until it reaches an advanced stage, making early detection challenging yet crucial for effective treatment. When applied to platelet-gene expression data from a published study, our machine-learning model could accurately discriminate between non-metastatic NSCLC cases and healthy samples. We further experimentally validated the panel on an in-house cohort of metastatic NSCLC patients and healthy controls via real-time quantitative Polymerase Chain Reaction (RT-qPCR) (AUC = 0.97). Model performance was boosted significantly after artificial data augmentation using the EigenSample method (AUC = 0.99). Lastly, we demonstrated the cancer-specificity of the proposed gene panel by benchmarking it on platelet transcriptomes from patients with Myocardial Infarction (MI).

3.2 Materials and Methods

3.2.1 Datasets

RNA sequencing dataset of TEPs used in this study was downloaded from GEO (Accession ID: GSE68086) [79]. We used 273 cancer samples with the following distribution: non-small-cell lung cancer (NSCLC): 59, colorectal cancer (CRC): 44, glioblastoma multiforme (GBM): 40, breast cancer: 38, pancreatic cancer: 33, hepatobiliary cancer (HBC): 5. There were 54 healthy samples. The dataset originally had 283 samples. We filtered out samples (n=10) with unknown labels, and low expression count. We also used TEP expression profiles from non-metastatic NSCLC cases (n=48) and healthy samples (n=344) from an independent study (GEO Accession ID: GSE89843) [136], as a test cohort. Gene expression profiles (raw read counts) were normalized using the TMM normalization method (edgeR package) [137].

In order to examine the gene panel's ability to classify early-stage cancer, we selected platelet RNA-seq samples consisting of 57 early locally advanced NSCLC patients (non-metastatic) and 377 healthy individuals from an independent study (GSE89843) [134].

To assess the specificity of our proposed gene-panel, we re-analyzed platelet transcriptomes from patients under Myocardial Infarction (MI) patients (GEO Accession ID: GSE109048) [138]. The dataset comprised of microarray gene expression profiles, obtained from 57 platelet samples with the following distribution: 19 ST-segment Elevation Myocardial Infarction (STEMI), 19 patients with Stable Coronary Artery Disease (SCAD), and 19 healthy donors. SCAD and STEMI are both phenotypically close conditions and cause changes in platelet gene expression relative to healthy controls.

3.2.2 Gene selection

In order to identify robust differentially enriched transcripts between healthy and cancer patients, we combined two independent statistical approaches. At first, we selected genes using on the Coefficient of Variance (CV). We selected the 1000 most variables genes, based on the CV score. We then overlaid the list on 1000 top genes sorted on the basis of differential expression test, conducted using Analysis of Variance (ANOVA). Of note, the differential expression analysis was performed across all 6 cancer types and healthy controls. A total of 11 overlapping genes were used to construct the panel. The workflow is outlined in Fig. 1.

3.2.3 Validation of the gene panel on RNA-Seq data

We used the selected genes to train classification models using three widely used techniques, namely Gradient Boosting Machines (GB) [139], Random Forest (RF) [129], and Linear Discriminant Analysis (LDA) [140]. To do this, we utilized the RNA-Seq read count data from a study by Best and colleagues [79]. As a benchmark, we considered comparing our predictions with ones obtained using 1000 genes that the authors proposed. We created 100 sets of 90-10 train-test stratified splits of the data for the area under the curve (AUC) measurements (Figure 3.2). We also checked the performance of the 11 genes using only NSCLC (n=59) and healthy samples (n=54).

To gauge the predictive power of the gene-panel for early cancer diagnosis, we chose

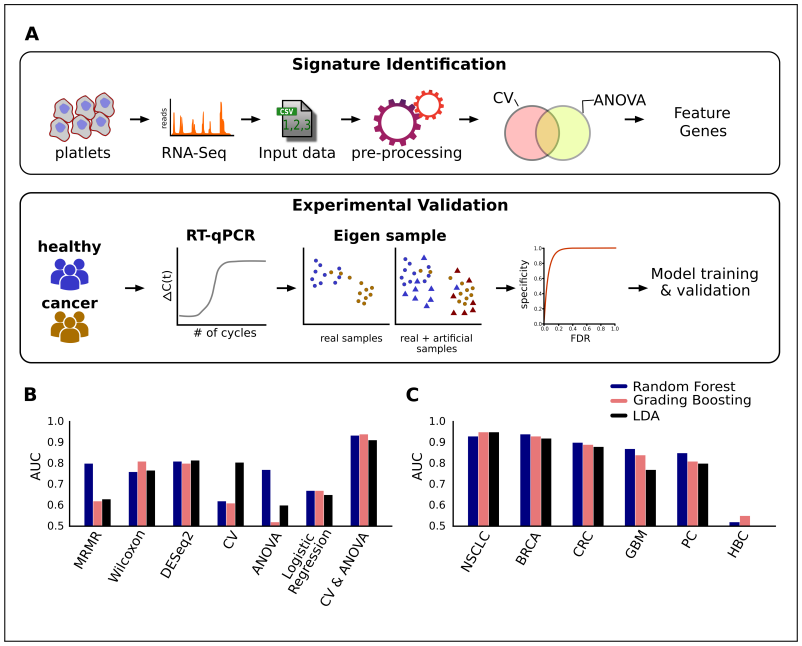

Figure 3.1: **Schematic representation of workflow.** **(A)** The upper panel is a schematic representation illustrating the underlying methodology implemented for the identification of the concise gene-panel utilizing RNA-seq data of tumour Educated Platelets (TEPs) (GSE68086). The lower panel represents the experimental design and the downstream statistical analysis employed in the validation of the inferred signature on a geographically distinct NSCLC patient cohort. **(B)** A comparison between different feature selection methods shows that a combination of Coefficient of Variation (CV) and Analysis of Variance (ANOVA) performs the best. **(C)** Classification accuracy across different cancer types.

non-metastatic NSCLC samples and healthy samples from GSE89843. To benchmark our findings against the reported values, we performed Leave-One-Out Cross-Validation (LOOCV) in tune with the methodology followed by Best et al. [134]. LOOCV for each classifier was performed 50 times with random seeds to measure the volatility of the models.

3.2.4 Clinical samples

Blood samples were collected from a total of 10 NSCLC patients and 7 healthy subjects (control) to train classifiers on data generated from the RT-qPCR experiment for validation purposes. We obtained ethical clearance from the Institute Ethics Committee at the All India Institute of Medical Sciences - New Delhi. All donors provided informed con-

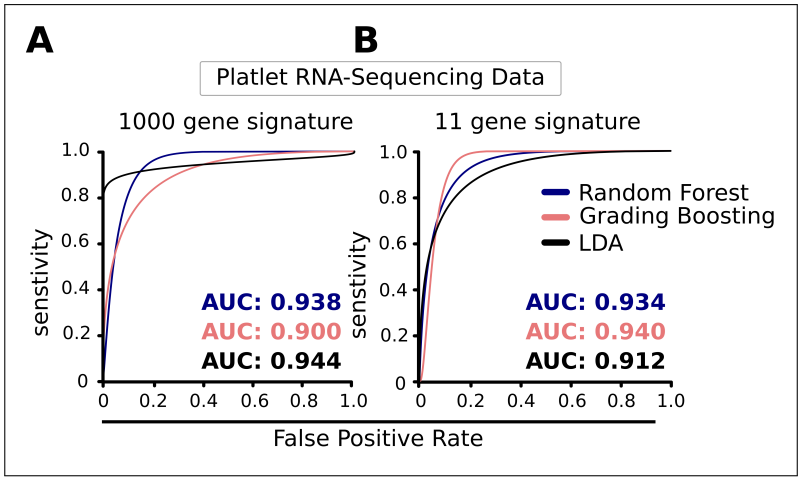

Figure 3.2: **Panel of 11 genes performs equivalent to panel of 1000 genes.** AUC (Area under the curve) plots representing the comparative performance of 1000 gene and 11 gene panels respectively on platelet transcriptomes from healthy and NSCLC patients. The predictive power of the gene-sets was evaluated using three widely used classification algorithms namely Gradient Boosting Machines (GB), Random Forest (RF), and Linear Discriminant Analysis (LDA).

sent before the collection of peripheral blood. 15 ml of peripheral blood was collected in a BD Vacutainer tube containing anticoagulant EDTA.

3.2.5 Platelets isolation from whole blood

The platelet-rich plasma (PRP) fraction was prepared by centrifugation of whole blood for 20 minutes at 120 x g at room temperature. The supernatant (PRP) was transferred into a fresh vial, and the red blood cell pellet was discarded after the first round of centrifugation. The platelets were enriched from PRP by centrifugation at 360 x g for 20 minutes at room temperature. The pellet representing platelets was washed with 1X Phosphate Buffered Saline (PBS) and centrifuged at 5000 rpm for 5 minutes. The PBS was discarded, and platelet pellet was re-suspended in $1ml$ TRI reagent® (SIGMA-Aldrich, USA) and stored at -80°C.

3.2.6 RNA isolation from platelets

Total RNA isolation was performed as per the manufacturer's recommendations (TRI reagent (SIGMA-Aldrich, USA)). Samples in the TRI reagent were thawed and mixed with 200 µl chloroform. After vigorous shaking, the samples were incubated for 15 minutes at room temperature, followed by centrifugation at 12000 x g for 15 minutes at 4°C. The aqueous layer was carefully transferred into fresh vials, and 500μl of isopropanol was added for RNA precipitation. After incubation for 10 minutes at room temperature, samples were centrifuged at 12000 x g for 10 minutes at 4°C. Next, we discarded the supernatant, and washed the RNA pellet twice with 75% ethanol, followed by centrifugation at 7500 x g for 5 minutes. After centrifugation, the RNA pellets were dried at room temperature and resuspended in 30 μl RNase-free water. RNA samples were quantitated using Nanodrop and stored at -80°C. cDNA synthesis was performed using the standard protocol as provided by the manufacturer cDNA kit (cat no. K1622, Thermo Fisher Scientific, USA). Briefly, the reaction mixture for cDNA synthesis was setup with 4μl 5X buffer, 2μl dNTPs, 1μl Random primer (RP), 1μl RiboLock (RL) and 1μl SuperScript Reverse Transcriptase and RNA sample in a total of 20μl volume.

3.2.7 Experimental validation of the gene panel using RT-qPCR

TaqMan gene expression assays (Applied Biosystems, California, USA) were used for expression studies of shortlisted gene candidates, namely *CD79B, CSDE1, IL-32, ITGA2B, LUC7L, NDUFAB1, RBM6, SKAP2, SS18L2, TRAF3IP3, and ZNF195.* Two reference genes (*ACTB* and *GAPDH*) were used as internal controls for downstream normalization steps. The reaction mix was prepared using 10μl Master mix, 1μl of gene expression assay, Nuclease-free water and cDNA sample per well.

3.2.8 Preprocessing of the RT-qPCR data

For the estimation of the relative gene expression, we used the comparative-Ct ($\Delta\Delta$Ct) method [141]. By using this approach, we first normalised our data using reference

genes and then calculated the relative expression differences for each gene (healthy vs cancer) by fold-change. Two different reference genes *ACTB* and *GAPDH*) were used for expression normalisation. We modelled our calculations and statistical analysis based on previously published examples [141, 142, 143].

3.2.9 EigenSample based artificial augmentation of the validation cohort

The EigenSample technique [144] was employed to augment the training data as sub-sampled from the entire set of RT-qPCR profiles. EigenSample simulates artificial data points in such a way tries to retain the original variance of the data. The pre-image of a new sample x^i is denoted by z^i and is obtained by solving the following optimization problem.

$$\underset{z^i, q^{i+}, q^{i-}}{\text{Minimize}} \quad \tfrac{1}{2}\|z^i\|^2 + C\sum_{j=1}^{k}(q_j^{i+} + q_j^{i-}) \tag{3.1}$$

s.t.

$$P \cdot z^i - q^{i+} \leq x^i + \epsilon \tag{3.2}$$

$$P \cdot z^i + q^{i-} \geq x^i - \epsilon \tag{3.3}$$

$$lb \leq z^i \leq ub \tag{3.4}$$

$$q^{i+}, q^{i-} \geq 0 \tag{3.5}$$

where ϵ is the approximation tolerance, and q^{i+} and q^{i-} are error variables. C is a hyper-parameter controlling the trade-off between the degree of approximation and norm of the solution vector $\|z^i\|$. A small value of C ($C \to 0$) will yield a minimum norm solution, while large C ($C \to \infty$) corresponds to the solution of a system of linear equations. Deploying machine-learning techniques on small sample sizes is difficult. Data augmentation is an important tool to increase the size of the labelled data and helps us to use the existing data more effectively [145, 146]. As such, we used EigenSample

to demonstrate that artificial augmentation of training data can improve the prediction outcomes.

3.2.10 Validation of the gene panel on RT-qPCR data

We used Leave-One-Out Cross-Validation (LOOCV) strategy to gauge the performance of various classifiers on RT-qPCR data. On every pass of LOOCV, we applied Eigen-Sample for training-data augmentation. For RF and GB classifiers, 50 random seeds were used to control their inherent stochasticity. The ROC plot was constructed while pooling predictions across these runs.

3.2.11 Exploring the co-regulatory network of the selected genes

To identify the potential transcription factors (TFs), regulating the empanelled genes, we extracted their putative promoter regions (1kb upstream of the transcriptional start site; TSS) using Eukaryotic Promoter Database [147]. Promoter sequences thus obtained were converted into FASTA format and were subjected to the Analysis of Motif Enrichment (AME) tool (a feature of the MEME suite), to discover common TF binding motifs [148]. For accurate inference of the common transcription factor binding sites (TFBSs) in the promoter sequences, we have utilized JASPER motif database [149], a reliable database harboring non-redundant transcription factor (TF)-binding profiles. Enrichment analysis of the common regulatory transcription factors was performed against randomly shuffled input sequences (control sequences). Fisher's exact test was used to report *p*-values. Differential expression of the TFs in the RNA-seq data [134] was calculated using edgeR [150]. Only NSCLC and healthy samples were selected from the dataset for the analysis. The enrichment of specific TF binding motifs in the promoter regions of our panel genes, coupled with their known roles in cancer, serves as a biological validation of our gene selection. It strengthens the argument that these genes are not just statistical markers but are biologically relevant to NSCLC. Our analysis using edgeR showed differential expression of these TFs in NSCLC samples

compared to healthy controls. This supports the hypothesis that the regulatory mechanisms involving these TFs are altered in NSCLC, affecting the expression of our panel genes.

3.3 Results

We analyzed a published, multi-cancer RNA-Seq data [79] with a set of 11 platelet genes (*CD79B, CSDE1, IL-32, ITGA2B, LUC7L, NDUFAB1, RBM6, SKAP2, SS18L2, TRAF3IP3, and ZNF195)* that enables accurate classification of cancer and healthy samples (refer Materials and Methods). We used Gradient Boosting Machines (GB), Random Forest (RF) and Linear Discriminant Analysis (LDA), three widely used classification methods to assess the potential of these genes in classifying cancer and healthy blood specimens. The best cross-validation accuracy was obtained using the GB classifier (AUC = 0.94), which matched the performance of the models that used 1000 variables, going by the recommendations of Best and colleagues ([79], Figure 3.2).

We performed independent experimental validation of the panel to assert two major reproducibility concerns. To this end, we used RT-qPCR to profile the expression of the selected 11 genes, on a cohort of 10 lung cancer patients (7 treat naive and 3 first-line chemotherapy) and 7 healthy controls (Figure 3.1A - lower panel). Gene expression trends, observed in our RT-qPCR data, were largely similar to that of the RNA-Seq study. Among the three classifiers, GB offered the highest accuracy (AUC = 0.97) (Figure 3.3(A)). RF and LDA offered AUC values of 0.87 and 0.74, respectively (Figure 3.3(A)).

We conducted a similar set of analyses on a distinct pathological condition, i.e. Myocardial Infarction (MI), where drastic shifts in the platelets transcriptome have been observed [138]. Since ST-segment Elevation Myocardial Infarction (STEMI) and Stable Coronary Artery Disease (SCAD) both cause perturbation in the platelet transcriptomes, samples with these conditions were grouped as one class (patients). The data, now having 2 classes (patient (n=38) vs healthy (n=19)), was then subjected to Leave-

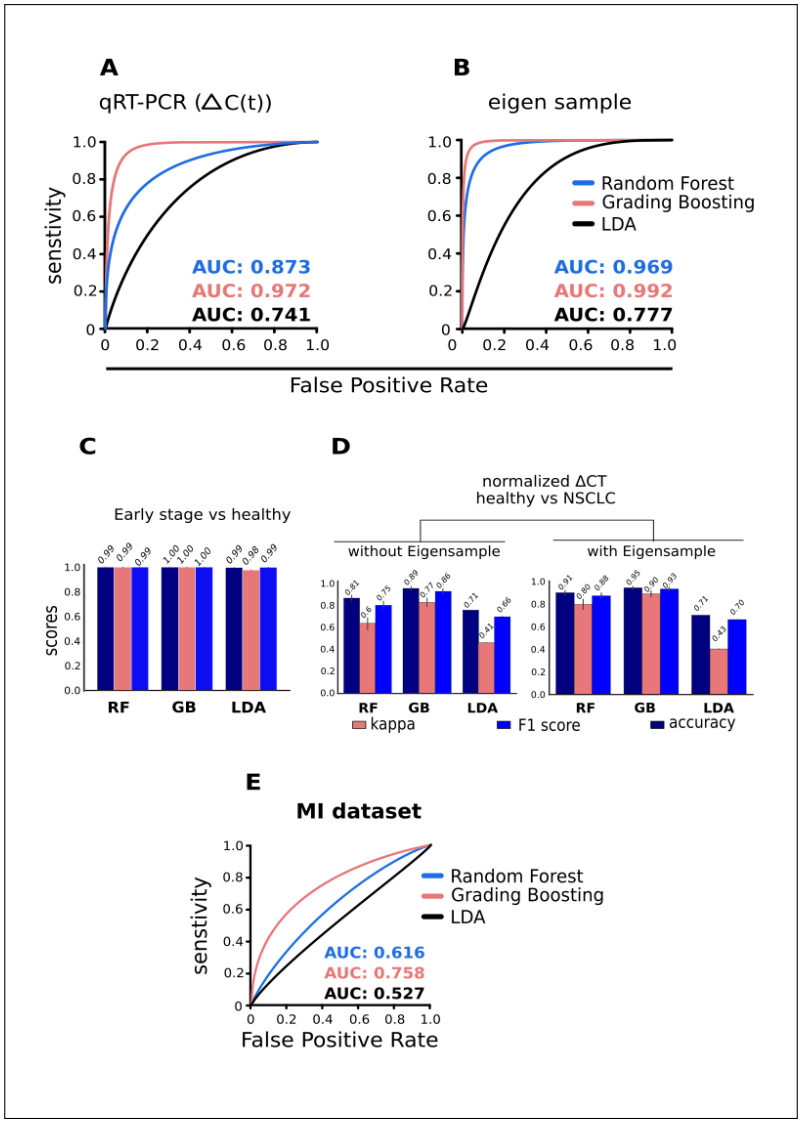

Figure 3.3: **Performances of three independent classifiers on early-stage vs healthy samples, MI samples and on RT-qPCR data (A)** AUC (Area under the curve) plot representing the performances of three independent classifiers i.e. Gradient Boosting Machines (GB), Random Forest (RF), and Linear Discriminant Analysis (LDA) in distinguishing tumour and healthy samples using Cq values of 11 genes from 10 NSCLC patients and 7 healthy controls. **(B)** AUC plot depicting the improvement in the classification accuracy by augmenting the data-points with artificial samples, using the EigenSample technique. **(C)** Classification performance based on the proposed 11 gene-panel the on TEP profiles of non-metastatic NSCLC patients and healthy controls from [134]. **(D)** Classifier performances on experimental data of 10 NSLC and 7 healthy samples. **(E)** Receiver Operating Characteristics (ROC) plot depicting the performances of three independent classifiers in distinguishing healthy and myocardial infarction episode samples using normalized intensity from platelets MicroArray dataset [138].

One-Out Cross-Validation (LOOCV) using 3 classifiers - RF, GB, LDA. Following the suite of NSCLC validation, RF and GB were run with 50 different seeds to estimate the

stochasticity of the models. As expected, our 11-gene signature failed to discriminate between the healthy and the diseased specimens under equivalent experimental settings, thereby suggesting the specificity of the signature towards the tumour datasets (Figure 3.3(E)). It should be noted that the overlap between early-stage NSCLC symptoms and other lung conditions such as COPD, also pose a heightened risk for developing NSCLC and it is indeed crucial to refine the specificity and sensitivity of TEP signatures. While the current study does not encompass analyses in this context, we acknowledge the importance of further studies that would calibrate TEP signatures against various pre-cancerous lung conditions. This approach would be instrumental in reducing the risk of over-diagnosis and ensuring timely and accurate treatment for those with NSCLC.

3.4 Discussion

Clinical implementation of non-invasive, liquid biopsy-driven molecular diagnostic methods often challenged due to the unavailability of robust, reliable and disease-specific molecular signatures. Though various statistical methods have been devised for the identification of reliable biomarkers from tumour-Educated Platelets (TEPs), their practical implementation in clinical diagnostics is still elusive due to a large number of proposed genes within the signature,. Therefore, to comprehensively address these limitations, we propose a novel albeit economically viable platelet-based classification assay for the early detection of cancer using an 11 gene signature. Importantly, we experimentally validated the identified signature with remarkable accuracy in geographically distinct patient-cohorts using a cross-platform assay. Taken together, by using a statistically-intensive, machine learning-based method, we propose a robust, reproducible and economically viable gene signature for tumour identification from the liquid biopsy.

This chapter demonstrates the application of computational analysis, particularly machine learning, to identify biologically significant markers for cancer diagnostics, with a focus on NSCLC. The development of the 11-gene panel, achieved through ma-

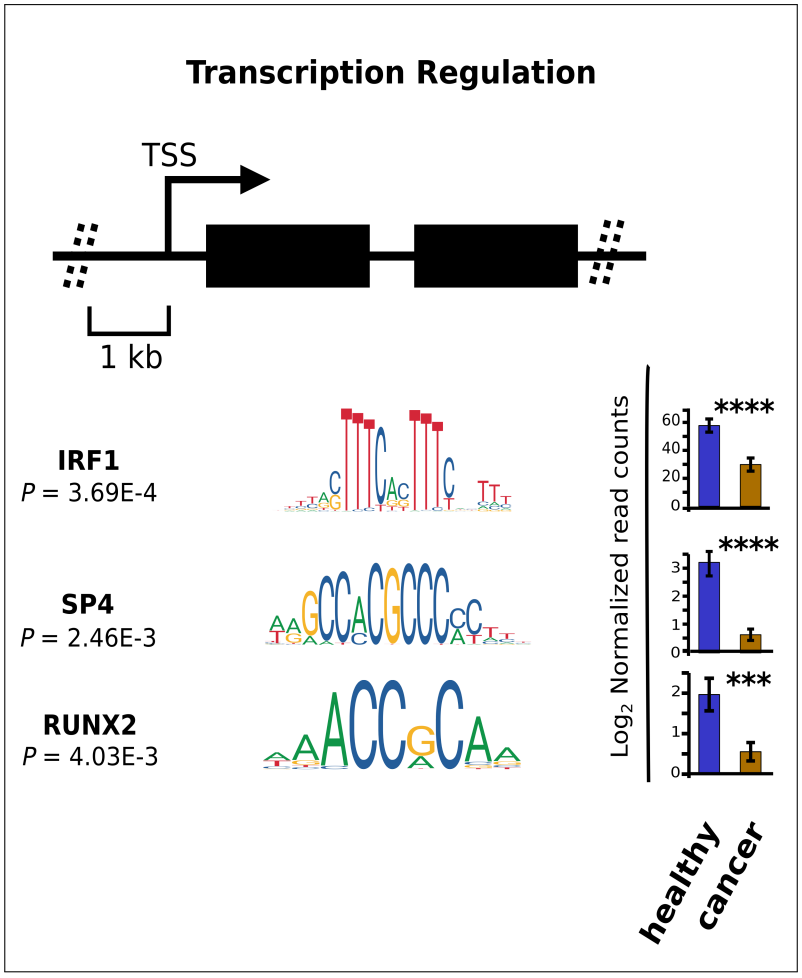

Figure 3.4: **Gene panels shares a regulatory circuit.** Graphical representation of the enriched transcription factor binding sites in the 1 kilobase upstream region (TSS=0) of 11 gene signature. *p*-value (FDR-corrected) represents the statistical power indicating a significant enrichment of the indicated motifs in the given region over shuffled control sequences. Bar graphs on the right represent normalized read-counts of the identified transcriptional factors between healthy and tumour samples. Asterisks represent *p*-value significance. *p*-value cutoff was set to 0.05. *, **, *** and **** represent the *p*-values of ≤ 0.05, ≤ 0.01, ≤ 0.001 and ≤ 0.0001 respectively.

chine learning techniques, represents a critical step in translating complex biological data into practical diagnostic tools.

Moving into the next chapter, we build upon this theme of patient-centric care. Next, we design machine learning strategies to guide the clinical management of cancer patients. Post diagnosis, the subsequent challenge is disease management. Our risk stratification model for Multiple Myeloma patients helps us cover the journey from diagnostics to optimized and personalized treatment plans.

CHAPTER 4

Developing a risk stratification model for Multiple Myeloma patients

4.1 Introduction and Motivation

We aim to identify patients who are at high risk of early relapse post stem cell transplantation. To do this we make use of only pre-transplant clinical variables. This shall aid doctors in deciding whether they should proceed with the stem-cell transplantation for a patient who is predicted to be at high risk of early relapse. Univariate and multivariate analysis is implemented to determine the risk groups of patients. Predicting risk of relapse for multiple myeloma patient if patients will go for autologous cell transplantation can help resolve many doubts. We analysed the data of 253 patients from 2005 to 2017, who underwent autologous stem cell transplantation (ASCT) at AIIMS, and received only novel agent regime for induction therapy.

Multiple myeloma (MM) is a cancer of plasma cells (a type of white blood cells) that is distinguished by clonal expansion of malignant plasma cells in the bone marrow and presence of monoclonal protein (M-protein) in blood and urine [151, 152]. MM is the second most common of all haematological cancer after non-Hodgkin lymphoma [153]. The main hypothesis is that the M-protein causes tumour growth by increasing cell proliferation and invasion. This leads to uncontrolled multiplication of cells which eventually results in metastasis. It is responsible for 15-20 percent of the deaths attributable to the haematological malignancies and about two percent of all cancer-related deaths [154]. Worldwide, it affects 1 to 5 per 100 thousand individuals per year with higher cases in the western countries [155, 156]. In a global, longitudinal study conducted concerning 32 cancer types, MM jumped from 23^{rd} place in 2005 to 21^{st} place in 2015 [155]. Racial disparity has been observed for both incidence and survival

in Multiple Myeloma. A US-based population study suggests that the occurrence of multiple myeloma in the black population is twice as compared to the caucasian population, while survival among the black population is significantly better compared to the caucasian population [157].

Introduction of novel agents over alkylating agents in induction therapy and high-dose chemotherapy followed by autologous stem cell transplantation (ASCT) considerably improved the survival of multiple myeloma patients in the past several years [158, 159, 160, 161, 162]. ASCT in the era of novel agents plays a crucial role in the management of younger MM patients. Patients receiving upfront ASCT have been found to have improved progression-free survival (PFS) and overall survival (OS) compared to patients receiving the conventional chemotherapy (CC) [163, 164, 165].

To address this issue, we analyzed clinical data of 253 multiple myeloma patients ((median age - 52 years, 166 males, 87 females), (between August, 2005 to December, 2016)), who were treated at the Department of Medical Oncology of All India Institute of Medical Sciences (AIIMS). We used Fast and Frugal Tree (FFT) for constructing a tree-based model for stratifying patients into either a high-risk or a low-risk group. The tree-based model included factors concerning 1. If the relapse occurs after remission, 2. response to induction therapy and 3. (pre-transplant) Glomerular Filtration Rate (GFR), which are commonly available prior to the transplant. Our 2-stage staging scheme yielded significantly distinct survival pattern between the risk groups both for progression free and overall survival.

4.2 Materials and Methods

4.2.1 Patients

Between April 1990 and December 2016, 347 patients with MM underwent ASCT at the Department of Medical Oncology of All India Institute of Medical Sciences (AIIMS). We collected clinical data of 347 patients. The drug regimen during induction

therapy has undergone significant changes since 1990's; with the recent trend being using novel agents. Written consent was obtained from all patients for the study. The study has been approved by the Institute of Ethics Committee, All India Institute of Medical Sciences(AIIMS) with the approval number: IEC-523/05.10.2018. For our study, we considered only those patients who were given the novel agents during induction therapy; since that is the latest mode of treatment. Such patients were 253 in number (median age: 52 years, range 29 – 68 years, 65.6% men).

4.2.2 Transplant protocol

Initially, all patients were reviewed in the weekly Bone Marrow (BM) Transplant Clinic in which the associated risks and benefits of bone marrow transplantation were explained to the patients and their family members. Written informed consent was obtained. Regimen-related toxicity was defined as per the Seattle criteria [166]. Pre-transplant evaluation included a detailed history, physical examination, staging according to the Durie and Salmon (DSS) [167] and the International Staging System (ISS) [168]. Details of previous treatment were recorded. The pre-transplant investigations included haemoglobin, total and differential count, renal and liver function tests, bone marrow examination, skeletal survey, and serum and urine electrophoresis, immune-fixation studies, serum β-2 microglobulin, and quantitative immunoglobulin levels.

4.2.3 Stem cells

The source of stem cells in most patients was granulocyte colony-stimulating factor (G-CSF) mobilized peripheral blood stem cells. Cyclophosphamide mobilized peripheral blood stem cells were used for stem cell harvesting in less than 10 patients. Even fewer patients had their stem cells harvested from bone marrow. The trypan blue dye exclusion test determined the viability of cells [169].

Induction therapy goes on for 4-5 months and usually consists of 4-6 cycles. The patients are treated with a combination of novel agents, e.g. immune modulators (thalido-

mide, lenalidomide), proteasome inhibitors (bortezomib) and dexamethasone, following which patients are treated with high dose melphalan [170].

4.2.4 Conditioning regimen

The myeloablative conditioning regimen consisted of melphalan dosage $150 - 225mg/m^2$ (218 patients, 86.2%) slow IV (intravenous) push on day 1 of transplantation followed by forced alkaline diuresis. Melphalan dosage of $\leq 150mg/m^2$ (35 patients, 13.8%) was given to patients with renal insufficiency (eGFR < $40ml/min/1.73m^2$, according to MDRD formula [171]) at the time of transplantation. With the change in melphalan dosage, no significant difference in the outcome of PFS and OS was observed. This is concurrent with previous literature [172].

Stem cells were transfused intravenously (i.v.) 24 hours after conditioning patients with high-dose of melphalan. $5\mu g/kg$ stem cells administered subcutaneously daily, including on day 0, 12 hours after stem cell infusion and onwards until engraftment. Patients were treated in isolation rooms and reverse barrier nursing was practised.

4.2.5 Data pre-processing

We used 39 variables (Table 4.1) from the clinical and laboratory data for the univariate analysis whereas 36 of them were used for the multivariate analysis. We ensured information related to these variables are typically available in the pre-transplant phase. We excluded ISS and DSS from the scope of the multivariate modeling since these are dependent on variables which already exist in our data. We excluded the variable depicting the number of line regimens since it's highly correlated with disease refractory status. According to our analyses it's inclusion gave rise model overfitting. Some of the variables had missing values (Figure 4.1), which were subjected to missing value imputation using an R package implementing MICE, a widely used algorithm for this purpose [173]. Categorical variables were transformed into numerical ones with the use of one hot encoding. This is essential for the machine learning based algorithms to

work.

Table 4.1: **Variabe Information**

Feature	Description
Age	Age of patient.
Gender	Gender of patient.
Height	Height of patient.
Weight	Weight of patient.
Body surface area (BSA)	Body surface area of patient.
Body Mass Index (BMI)	Body Mass Index of patient.
Hb	Hemoglobin level (g/dl) in blood when patient comes for first time.
Creatinine	Creatinine level in blood when patient comes for the first time. High level creatinine indicates malfunctioning of kidney. 3.0.19 Discussion
Albumin	Albumin level in blood when patient comes for the first time. Lower serum albumin levels in multiple myeloma patients are associated with clinical factors reflecting disease severity.
Immunoglobulin type	Immunoglobulin isotype in blood whether IgG-Kappa, IgG-Lambda, IgA-Kappa, IgA-Lambda, Kappa, Lambda, IgMK, nonsecretory, IgD, IgM.
Beta-2-microglobulin (B2M)	The beta-2-microglobulin level in blood when the patient comes for the first time. B2M is the most powerful prognostic predictor of multiple myeloma.Lower serum albumin levels in multiple myeloma patients are associated with clinical factors reflecting disease severity. [168]
Serum Calcium	Serum calcium level in blood when the patient comes for the first time. High calcium level in serum is related to advance multiple myeloma. It shows osteoclast.
Extramedullary disease (EMD)	Whether patient have Extramedullary disease or not. EMD is having myeloma tumours outside the bone marrow in soft tissues of body.
Response to induction therapy	How patients responded to induction therapy, whether they showed complete response, very good partial response or partial response, Stable disease, No respose.
Absolute Neutrophil Count (ANC)	Neutrophil count when patient comes for the first time.
Absolute lymphocyte count (ALC)	Lymphocyte count when patient comes for the first time.
Glomerular filtration Rate (GFR)	Glomerular Filtration Rate when the patient comes for the first time. It represents renal condition. Lesser the GFR associated with clinical factors reflecting disease severity.
GFR Score	GFR score is the categorical variable which shows how GFR changed over the induction therapy.
GFR Grade	Categorised GFR level: "1=>=90ml, 2= >=60 to 89, 3=>=30 to 59, 4=15-29, 5-<15ml".
Primary (If disease is refractory)	Patients categorised into two categories. 1= patients who received treatment then went on to receive stem cell transplant, 2= patients relapsed after initial treatment than received salvage therapy then received stem cell transplant. Thus, it tell us whether disease is refractory.
Stem cell Harvest site	Stem cell harvest site, whether harvest site is mobilized peripheral blood stem cells or bone marrow.
BMPC	Count of Bone Marrow Plasma cell while harvesting.
Hypertension	Whether patient have hypertension or not.
Diabetic	Whether patient have diabetes or not.
Line of induction therapy	How many different types of regimens given in induction therapy. One line consist 3 to 6 cycles (typically 4). One line means only one type of regimen is given in all cycles. If a new combination of drugs is given than previous, then it will consider in second line and so on.
Number of cycles in 1st line treatment	Regimens given in one cycle is combination of different drugs. Number of cycles in one line of induction therapy.
Radiation Therapy	Whether patient received radiation therapy or not. Radiation therapy is given in the preparation of stem cells to kill myeloma cells.
Erythropoietin Treatment (Epo)	Whether patient received Erythropoietin treatment or not. Epo treatment is associated with improved immunological functions.
Dialysis	Whether patient receive dialysis or not . Renal failure is the main reason of relapse in multiple myeloma patients. [174, 175]
ISS	International staging system calculate three stages of multiple-myeloma by measuring serum albumin and beta-2-microglobulin level . stage I associated with less severity and stage III associated with highest severity [168].
DSS	Durie-Salmon Staging System calculate staging of multiple myeloma patients by measuring blood calcium level, hemoglobin level, M protein level and kidney function. Higher the DSS stage associated with higher severity [167].
Platelet count	No of platelets when patient comes for the first time.
Monoclonal protein level	Monoclonal protein level in the blood when patient comes for the first time. Higher the level of monoclonal protein associated with severity of disease.
Symptom duration	Duration of symptoms in months.
Pre-transplant creatinine level	Creatine level after induction therapy and before autologous stem cell transplantation.
Pre-transplant GFR	Glomerular filtration rate after induction therapy and before autologous stem cell transplantation.
Pre-transplant M-protein level	Serum M protein level after induction therapy and before autologous stem cell transplantation.
Melphalan dose	Dose of melphalan given as the conditioning regimen after collecting stem cells.
CD34 cells number	Number of CD34 cells in stem cell harvesting. It indicates the purity of stem cells in the sample.

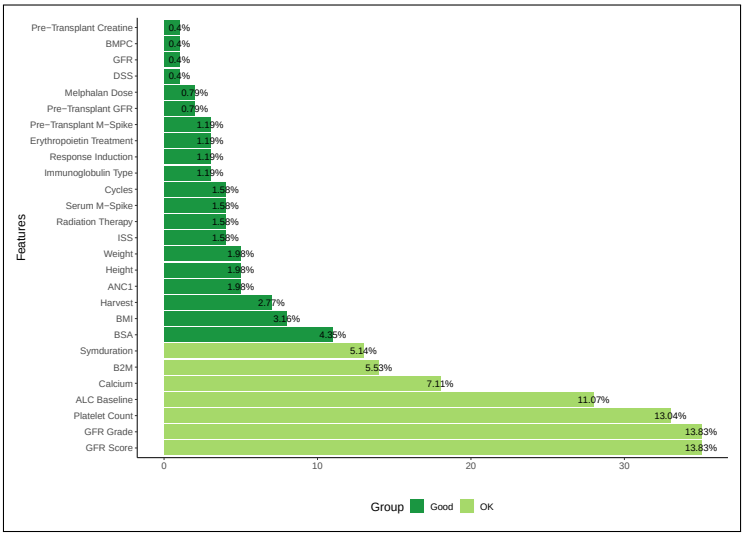

Figure 4.1: **Percentage of missing values**

4.2.6 Univariate analysis

Associations of the individual factors w.r.t. OS and PFS were analyzed using the widely used Kaplan Meier's survival analysis technique [176]. Categorical variables were grouped by categories, whereas the numerical variables (23 out of 39) were subjected to univariate K-Means [177] for exploring groups. For simplicity, the number of clusters was set to 2 for each case. A cut-off value was generated based on the highest observed value of the cluster comprising the smaller values. If the highest observed value is C, the associated ranges are $\leq C$ and $> C$.

4.2.7 Multivariate analysis

Typically, variables in combination hold promise for a more nuanced predictive model. Predictive modeling involves training of the model, followed by validation. When the sample size is small, taking out data-points for validation turns out to be detrimental as it weakens the model training. On the flip side, training a model on the entire data is usually suspect for model overfitting.

We bypassed this problem by developing a two-pronged learning approach. We first grouped the patients using spectral clustering [178]. For this, we constructed an adjacency matrix spanning the data points (patients) by computing the Hamming distance of each pair of points. Continuous variables were considered in their binary form for the distance calculation. Principal Component Analysis (PCA) was performed on the distance matrix. Principal Components entailing 95% of the Eigen energy were subjected to spectral clustering. Two clusters thus obtained showed distinct survival patterns both for OS and PFS (Figure 4.2, Figure 4.3).

We treated the clusters as high-risk and low-risk groups. To aid clinical decision making, we fitted a Fast and Frugal Tree (FFT) [179] for accurate prediction of risk groups. An FFT is a simpler version of a decision tree [180]. The most striking feature of FFT is that unlike decision tree it is usually simple enough for a human mind to memorize. FFTs have been shown to perform competitively with random forest [180].

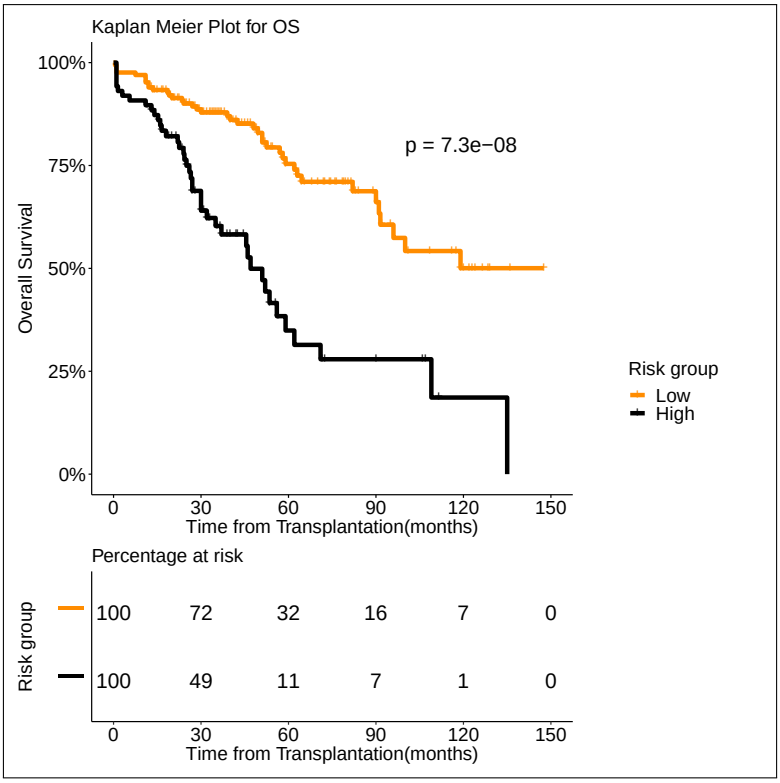

Figure 4.2: **Overall survival (OS)** in 253 patients with multiple myeloma stratified by spectral clustering. Median OS was more than 90 months for low risk group (number of patients = 166, events = 40) (shown in orange), whereas it was 47 months for high risk group (number of patients = 87, events = 42) (shown in black).

FFT, being conspicuously simple, does not warrant overfitting. Therefore, we refrained from independent validation of tree performance. On the first pass, we trained a 2-class FFT, while treating the cluster identities as the labels for the patients under study. We then subjected the samples to the trained FFT to re-calibrate the labels. Contrasting the predicted labels with the ground truth resulted in 84.9% accuracy. As evident from the accuracy (84.9%), the FFT managed to model the clusters. In fact, the slight modifications of the labels due to re-calibration caused an increase in the median survival of the low-risk group in case of PFS (91 months instead of 74 months), while the median survival of the high-risk group remained unaltered (24 months).

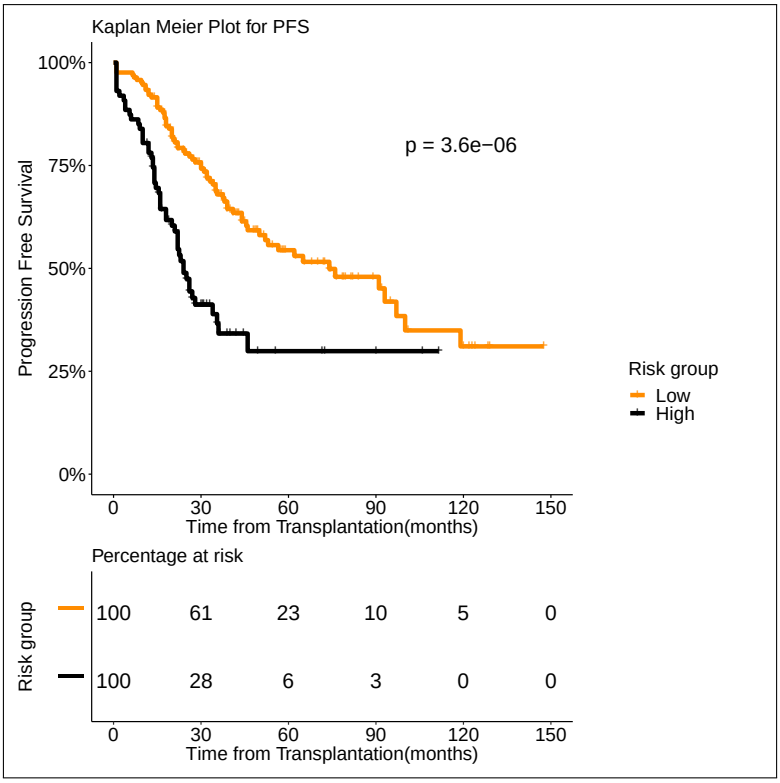

Figure 4.3: **Progression Free survival (PFS)** in 253 patients with multiple myeloma stratified by spectral clustering. Median PFS was 74 months for low risk group (number of patients = 166, events = 70) (shown in orange), whereas it was 24 months for high risk group (number of patients = 87, events = 50) (shown in black).

4.3 Results

Multivariate analysis using spectral clustering gave us two groups; which were significantly different; as confirmed by Kaplan Meier and FFT classification analysis. (Figure 4.4) depicts year wise distribution of patient frequencies. Patients were administered various drug combinations during induction therapy. The drug regimen for induction therapy has undergone significant changes since the '90s, usage of novel agents being the current trend. VAD regimen (Vincristine (V), Doxorubicin (A), Peroral Dexamethasone (D)) was administered to 72 patients (between April 1990 and March 2005). Alkylating agents were given to only 22 patients (between July 1997 and March 2014). As expected, novel agents yielded improved survival as compared to VAD and Alkylating agents. 253 patients were treated with novel agents from August 2005 to December 2016. For our study, we considered only the 253 patients who were treated with various

novel agents (Table 4.3.) during induction therapy, since that has been the most prevalent mode of treatment during the last decade. No significant difference was observed in the survival trends across the novel-agents (Figure 4.5). No patient was lost in follow up. Follow up was done till 30^{th} November 2017 (date of censor). For patients treated with novel agents, 8 out of 253 had undergone dialysis. Post-transplant, only one of these 8 patients had undergone elective dialysis. The patients subsequently underwent renal transplant as well and continued to be disease-free for more than 2 years. Some important patient characteristics are shown in Table 4.2.

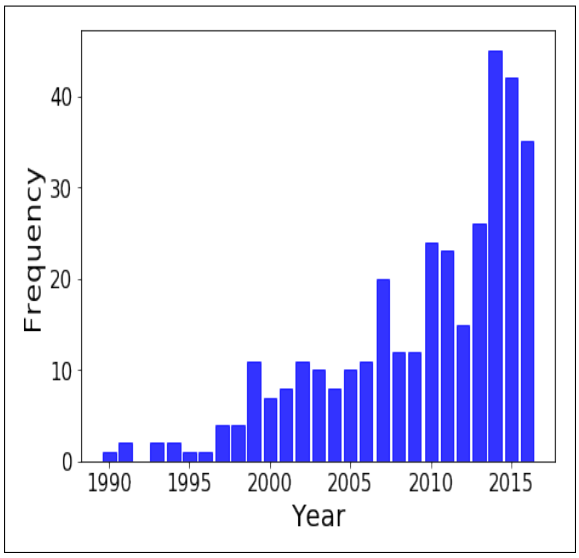

Figure 4.4: **Patient distribution over the years.**

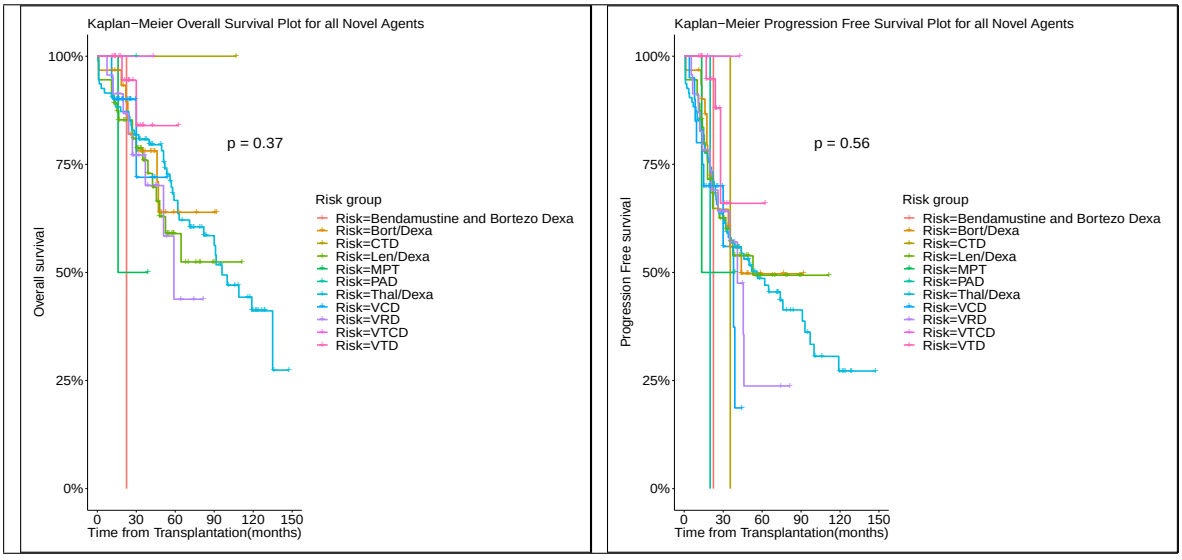

Figure 4.5: **Comparing OS and PFS of various Novel agents.**

Table 4.2: **Patient characteristics**

characteristics	Number of patients out of 253
Median Age	52
Gender	166 males, 87 females
Relapsed after remission, then transplant	69
Diabetic	107
Presence of Extramedullary Disease	57
Renal Condition	
Required Dialysis	8
ISS Stage during diagnosis	
Stage-I	67
Stage-II	94
Stage-III	92
Line of induction Therapy	
One line	174
Two line	54
Three line	19
>Three line	6
Immonoglobulin Type	
IgG-Kappa	101
IgG-Lambda	43
IgA-Kappa	26
IgA-Lambda	15
Kappa	45
Lambda	23

Table 4.3: **Table showing frequency of novel agents**

Novel Agent type	Frequency
Thal/Dexa	94
Len/Dexa	55
Bort/Dexa	31
Bortezomib/Thalidomide/Dexamethasone (VTD)	23
Bortezomib/Lenalidomide/Dexamethasone (VRD)	23
Bortezomib/Cyclophosphamide/Dexamethasone (VCD)	20
Cyclophosphamide/Thalidomide/Dexamethasone (CTD)	2
Bortezomib/Thalidomide/Cyclophosphamide/Dexamethasone (VTCD)	2
Melphalan/Prednisone/Thalidomide (MPT)	1
Bendamustine and Bortezo dexa	1
Bortezomib/Low-dose dexamethasone/Pegylated liposomal doxorubicin (PAD)	1

4.3.1 Factors affecting response to transplant

We performed Kaplan Meier's survival analysis for the individual factors to determine the ones that have prognostic value. Out of a total of 39 factors (Table 4.1), 23 were numerical (pre-transplant M-protein level, pre-transplant GFR etc.). We grouped patients

based on each numerical feature using univariate K-means (Methods). We tracked both overall and progression-free survival for each of the factors. Factors that displayed remarkable prognostic value for both OS and PFS were: 1. If patient relapsed following remission ($P < 0.001$ for both OS and PFS), 2. the number of regimens used pretransplant ($P < 0.001$ for both OS and PFS), 3. serum albumin level ($P < 0.001$ for both OS and PFS) and pre-transplant M-protein level ($P = 0.0018$ for PFS and $P = 0.002$ for OS). ISS staging appeared minimally predictive for PFS. Important outcomes of the univariate analyses are captured in (Table 4.4).

Table 4.4: **Prognostic power of the important individual factors**

Factor		**n**	**Progression Free Survival**					**Overall Survival**				
			Events	**Median**	**P Value**			**Events**	**Median**	**P Value**		
Induction line	**1**	174	69	76	<0.001			41	135	<0.001		
	>1	79	51	24				41	47			
Albumin	**<=3.5**	103	58	34	<0.001			45	56	<0.001		
	>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			
Pretranplant 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

4.3.2 Multi-factor survival modeling

We found multiple variables to have an independent association with survival. Moreover, single variable risk stratification is of limited use for its restrictive nature. For instance, it may so turn out that a fraction of first line patients relapse soon after the graft. If we create a rather simplistic single factor staging scheme just based on the relapse (after remission) status, it may under-predict for those at risk. For multi-factor modeling exercise, a major hindrance is small sample size. Commonly used methods such as survival tree [181] requires a large number of samples to produce a meaningful model. For instance, the International Staging System (ISS) was built on clinical and

laboratory data of about 10,000 patients [168].

We first examined the heterogeneity in the patient population using spectral clustering (see Methods). All 39 pre-transplant variables were used for this. We obtained two clusters that showed a stark difference in survival pattern both for PFS ($P < 0.001$) and OS ($P < 0.001$). See Figure 4.2, Figure 4.3 for the associated Kaplan Meier analysis. We marked the patient-groups mirrored by the clusters as high-risk and low-risk depending on their survival trend. The high-risk group consisted of 34% of the patients with a median progression-free survival of 24 months. On the contrary, the low-risk group consisted of 66% of the patients with a median progression-free survival of 74 months (Figure 4.3). While clusters are useful to unravel patient heterogeneity, they don't augment clinical decision making. To this end, we used a novel iterative approach for constructing a Fast and Frugal Tree (FFT) that effectively models the clusters (see Methods). The tree is meant for mapping any patient to one of the risk groups depending on their characteristics.

FFT based modeling offered a simple, 3-factor decision tree that predicts the risk category of a patient. It's similar to a staging scheme. Variables elected by the final FFT included - 1. If patient relapsed following remission, 2. response to induction and 3. pre-transplant GFR (Figure 4.6).

Subjecting patients to the FFT showed better discrimination in survival patterns across the re-calibrated high-risk and low-risk groups (Methods). While the median progression-free survival of the high-risk group remained unchanged (24 months), for the low-risk group, we obtained a median survival of 91 months (see Figure 4.7, Figure 4.8) for the KM analyses for OS and PFS). Notably, we found the risk-groups to have partial concordance with the variables having independent prognostic value.

We excluded ISS and DSS from the scope of the multi-variate modeling since these are dependent on variables which already exist in our data. We excluded the variable depicting the number of induction lines since it's highly correlated with disease relapse (after remission) status. Its inclusion gives rise model overfitting.

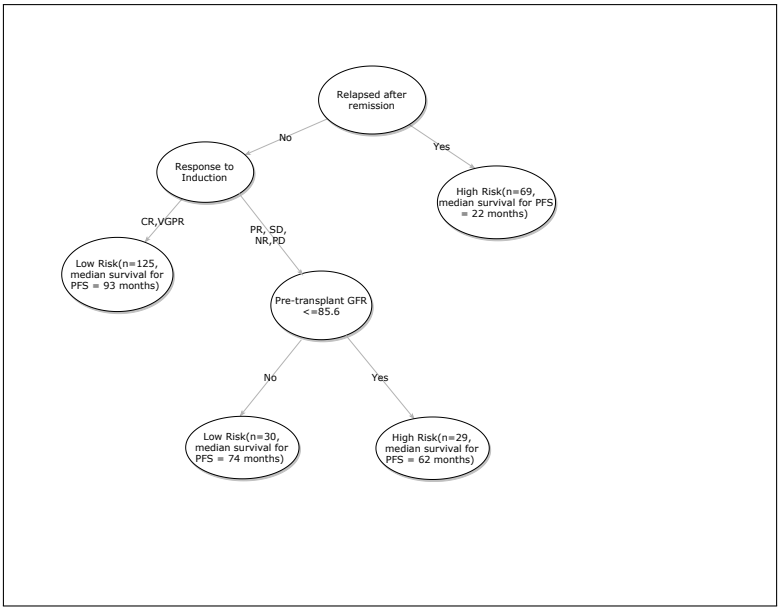

Figure 4.6: **Fast-and-frugal tree based staging scheme for patients undergoing ASCT.** CR=Complete Response, VGPR = Very Good Partial Response, PR = Partial Response, NR = No Response, SD = Stable Disease, PD = Progressive disease.

4.3.3 Prognostic value of alternative staging systems

We evaluated the prognostic value of the existing, widely practiced staging systems — Durie Salmon Staging (DSS) [167] and International Staging Systems (ISS) [168]. DSS relies on haemoglobin concentration, level of blood calcium, the presence of bone lesions, M protein level in urine and blood and kidney function level to predict the extent of the disease. ISS, on the other hand, uses albumin and Beta-2-microglobulin levels for staging patients with MM. One must note that DSS and ISS are not meant for predicting the outcome of stem cell rescue. We found DSS to be an extremely weak predictor of the ASCT outcome. We observed some association across ISS-I - ISS-II(P = 0.0556), and ISS-I - ISS-III (P = 0.0042) in case of OS. For PFS, ISS staging turned out to be a weak predictor (ISS-I - ISS-II (P = 0.2), and ISS-I - ISS-III (P = 0.15)).

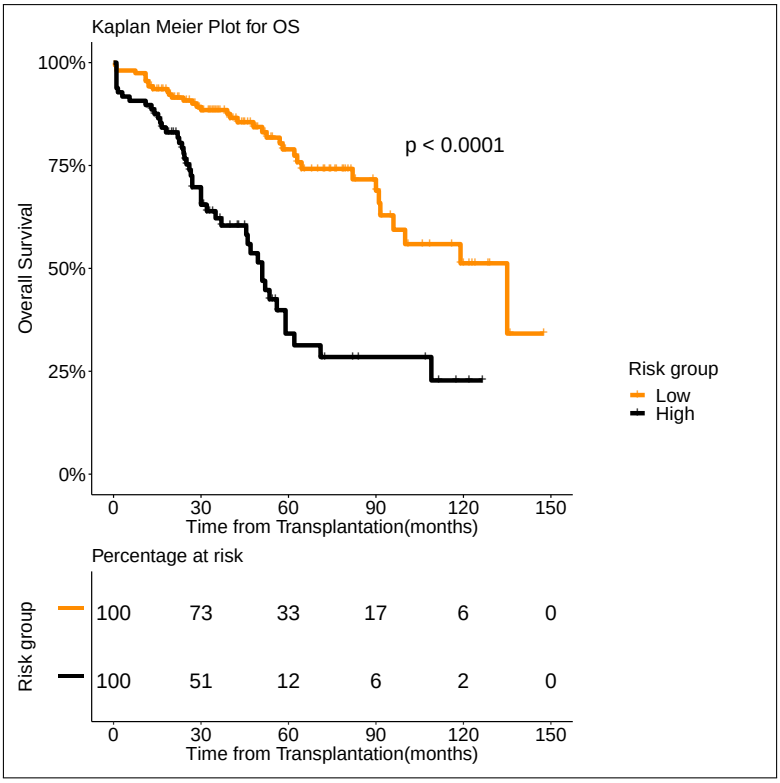

Figure 4.7: **Overall survival (OS)** in 253 patients with multiple myeloma stratified by FFT rules. Median OS was 135 months for low risk group (number of patients = 156, events = 36) (shown in orange), whereas it was 51 months for high risk group (number of patients = 97, events = 58) (shown in black).

4.4 Discussion

Over the last decade autologous stem cell transplantation (ASCT) has emerged as the standard of care in the management of Multiple Myeloma (MM). However, the cases of early relapse (within 36 months) after the stem cell rescue remains a significant challenge. For a lot of practical purposes, it is crucial to identify whether a patient undergoing ASCT falls into the high-risk group (likely to relapse within 36 months) or a low risk one. Our analysis showed that existing MM staging systems (International Staging System or ISS and Durie Salmon Staging or DSS) are not sufficient to discriminate between the risk groups significantly. We aim to identify patients who are at high risk of early relapse post stem cell transplantation using pre-transplant clinical variables. This shall aid physicians in deciding whether they should proceed with the stem-cell transplantation for a patient who is predicted to be at high risk of early relapse.

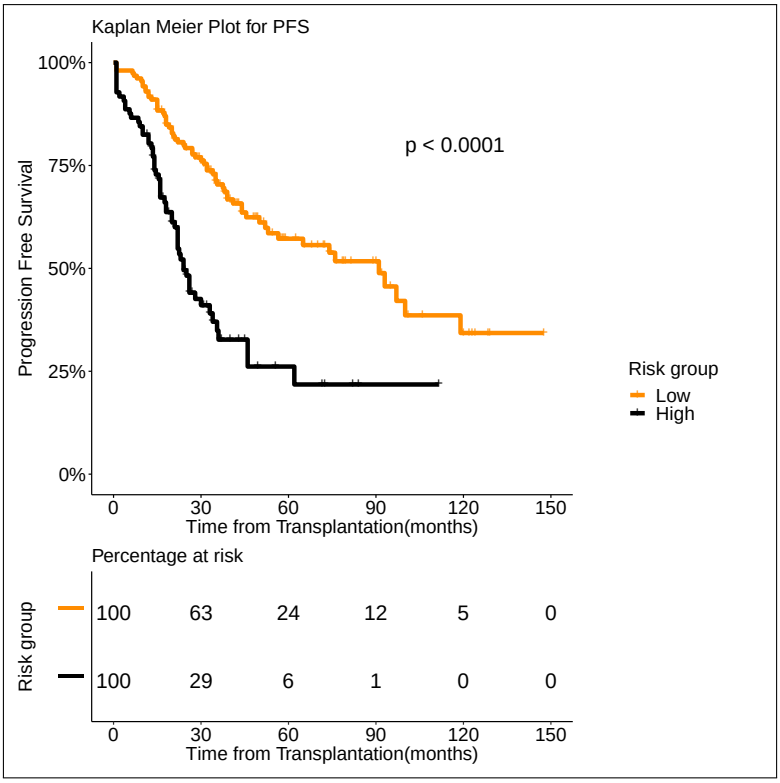

Figure 4.8: **Progression Free survival (PFS)** in 253 patients with multiple myeloma stratified by FFT rules. Median PFS was 91 months for low risk group (number of patients = 156, events = 62) (shown in orange), whereas it was 24 months for high risk group (number of patients = 97, events = 58) (shown in black).

Our goal is to empower doctors with a statistics driven opinion on whether a patient is at risk of early relapse. In such cases, doctors may try and find alternate solutions. We consider it to be a case of early-relapse, when the patient relapses within 36 months (3 years). Such patients are termed as High-Risk. Another case where our study may be helpful is to convince patients to undergo ASCT, in case they opt-out of stem-cell transplantation after induction therapy. Only 20%-30% of patients undergo ASCT after induction therapy. Physicians shall be empowered with a second statistics-driven opinion to convince patients who wish to opt-out of ASCT to not do so, in case they fall in Low-Risk strata. Studies have shown that multiple-myeloma patients always have a better chance of survival in case they go for stem-cell transplantation[164].

We used clinical and lab data of 253 patients who have been treated with novel agents and undergone ASCT at AIIMS between 2005 and 2016. Due to the small sample size, we developed a new machine learning approach that's minimally susceptible to

the problem of model overfitting. We showed that a simple, 3-variable (if relapse occurs after remission, response to induction and (pre transplant) GFR) decision tree can serve as a staging scheme that maps each patient to one of the two (high and low) risk groups, with markedly distinct survival patterns for both overall and progression-free survival.

As per the proposed tree based model patients with relapsed disease (following remission) were predicted under the high-risk category. Median PFS of relapsed patients is 22 months compared to that of first-line patients which is 76 months (Table 4.4). Patients with relapsed MM do better with ASCT but relative to other patients (non-relapse) their survival is poor [182, 183]. The model correctly identifies the state of relapse as the key factor for risk prediction. As previous literature suggests, 90% of the patients exhibiting complete response to induction therapy, also exhibit complete response to ASCT. In case of very good partial response during induction, the corresponding complete response is 72% [170]. The patients who do not show complete response and very good partial response are likely to relapse quicker [184]. The model correctly identifies this variable as the next important factor for relapse prediction. Renal functioning is an important factor for multiple myeloma patients since renal insufficiency is positively correlated with increased mortality [185]. GFR grade is defined as $\geq 90, \geq 60-89, \geq 30-59, 15-29, \leq 15ml/min/1.73m^2$ as per the standard criteria [186]. The model predicted cutoff of 85.6 closely approximated the recommended cutoff for stage 1 i.e., GFR $\geq 90ml/min/1.73m^2$.

Many studies suggest age as an important predictive factor for ASCT outcome [187, 170]. We observed that patients regardless of age, appear to benefit from ASCT. This has been seconded by previous studies [188], as well as in our data. The median age of patients treated with novel agents is 52 years. No significant differences (P = 0.31 for PFS and 0.36 for OS) were observed between the two groups (<= 52 yrs and >52 yrs). As a result, age has not been picked up by the decision tree among the top influencers.

The model we obtained also highlights the importance of multivariate analysis. Pre-transplant GFR, independently, did not emerge as an important prognostic factor (Table 4.4). However, when combined with the relapse (following remission) status and re-

sponse to induction therapy it led to a more nuanced stratification of the patients into the risk categories. Despite no stark difference in the median PFS (high-Risk - 62 months; low-risk - 74 months), patients subjected to the GFR mediated bifurcation showed a significant difference in the rates of 5-year survival rate (37% vs. 14%).

Due to data paucity, we did not apply excessive inclusion or exclusion criteria besides considering only those patients who were treated with novel agents. Our multivariate model discerned the variable outcomes between newly diagnosed and relapsed (after remission) patient groups using a limited number of pre-transplant clinical variables. This also serves as a testimony for the model's inherent ability to accurately predict graft outcome for diverse patient strata. The proposed tree based model labels relapse after remission cases as 'High-risk' (Figure 4.6).

To test the ubiquity of the newly proposed staging scheme we employed additional data of patients, treated with VAD and alkylating agents. Application of our 3-factor rule sets stratified the mixed pool of patients into the high and low-risk categories. Kaplan Meier survival analysis yielded distinct (overall and progression-free) survival patterns ($P < 0.0001$ in both cases), following the trend observed on the patients treated with novel agents (Figure 4.9, Figure 4.10).

A limitation of the study is the unavailability of cytogenetic/FISH data which is incorporated in the revised ISS system [189]. The patient data is collected over a long period of time (2005-2017) and we did not have cytogenetic /FISH data for the initial period (till 2011). Another shortcoming of the current study is sample paucity. We plan to perform a multi-centre follow-up study to ascertain the integrity of our staging scheme.

In this chapter, we extend the computational approaches from dissecting the molecular biology of cancer (Chapter 2) and diagnostics (Chapter 3) to the realm of prognostics in cancer care. The predictive model for risk of relapse prediction integrates the themes of the previous chapters, applying computational methods to address a different but equally crucial aspect of cancer treatment. This chapter culminates the thesis by showcasing the full potential of computational tools in cancer research, from ana-

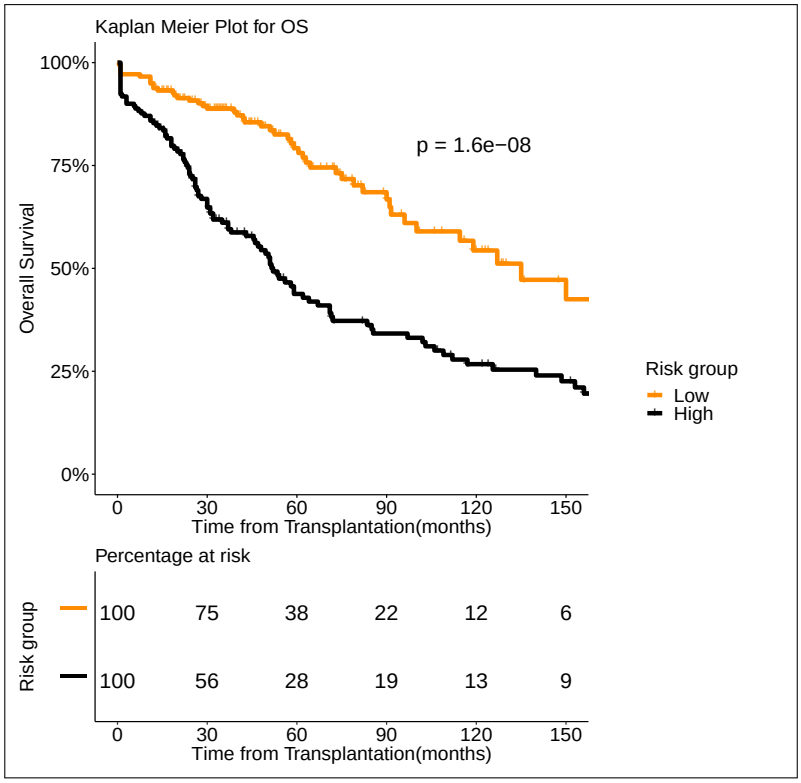

Figure 4.9: **Overall Survival (OS)** in all patients with multiple myeloma stratified by FFT rules. Median OS was more than 135 month for low risk group (shown in orange), whereas it was 52 months for high risk group (shown in black).

lyzing heterogeneity to identifying diagnostic markers and, finally, guiding the clinical management of cancer patients.

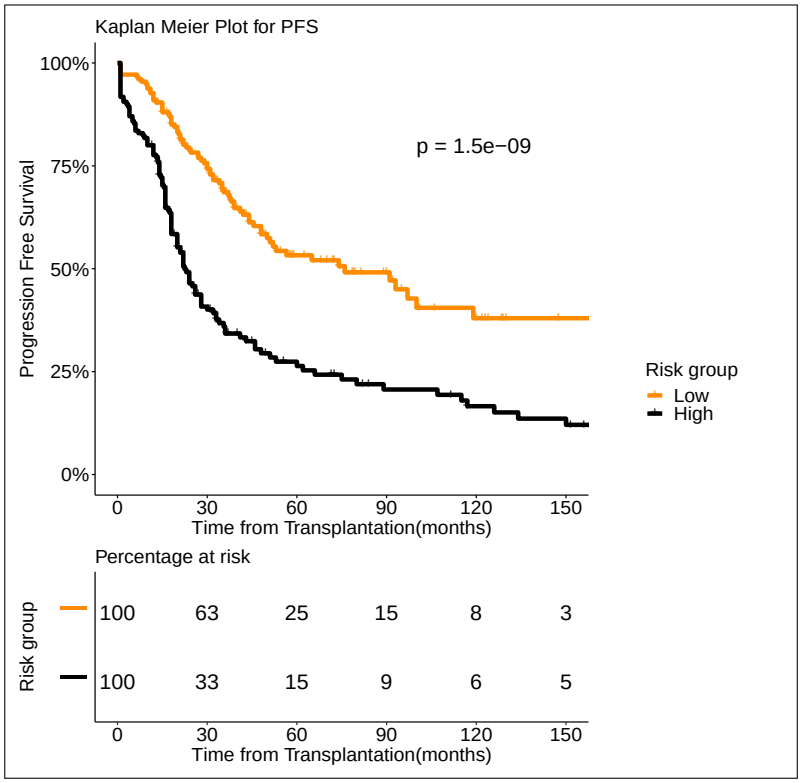

Figure 4.10: **Progression Free Survival (PFS)** in all patients with multiple myeloma stratified by FFT rules. Progression Free median OS is 76 month for low risk group (shown in orange), whereas it is 22.5 months for high risk group (shown in black).

CHAPTER 5

Conclusions

This thesis focuses on developing machine learning based strategies for biomarker discovery, affordable molecular diagnostics and personalised medicine in cancer. Our work focuses on interpretable and easy-to-implement models to solve important public health and computational biology problems. This thesis effectively integrates machine learning and statistics to derive clinical and genomics data insights. We demonstrate how machine learning can help clinical decision support systems in real-world scenarios with a multitude of data types.

Getting the laboratory discovery into the clinic comes with a fair share of challenges. Laboratory discoveries are often not useful as treatments because experiments in laboratories are conducted in controlled environments, with many constraints in place. In the real world, data gets very messy. At finer resolutions, the data is sparse, and it can suffer from high levels of noise. Some other challenges are: large dimensions, lack of methods that scale, missing information. To bring solutions from bench to bedside, some other facets need consideration, like- experimental validation, affordability, reproducibility on geographically distinct populations, and disease specificity. We have addressed some of these real-world challenges in our work. This thesis demonstrates how we can leverage machine learning in translational cancer research to make the journey from the laboratory to the clinic to improve patient care. This thesis provides solutions that can translate in the real world across multiple areas, such as patient stratification, cost-effective biomarker analysis, and therapy recommendations, which in turn help reduce the burden of cancer on the patient. Overall, this research is a significant feat to address the broader issues of accessibility, reproducibility, and disease specificity.

5.1 Summary of contribution

In this section, we summarise the chapters, giving a comprehensive view of the thesis.

5.1.1 Finding influential genes from embeddings of non-linear dimension reduction techniques

Dimensionality reduction is often used to visualize high-dimensional and complex data. Dimension reduction techniques are routinely used to facilitate downstream analyses and visualize high-dimensional datasets. Such methods replace several thousands of genes with a small number of latent variables while preserving important patterns in datasets. In the field of single-cell transcriptomics, Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation Projection (UMAP) are popular. Following dimension reduction, an obvious question to ask is, "What are the key genes?". For PCA, often genes with the highest loadings in the first few principal components are considered to be the most influential. t-SNE and UMAP are non-linear dimension reduction techniques that can identify hidden data structures and expose natural clusters and non-linear variations along the dimensions. These techniques are often used to visualize high-dimensional data, but they do not have an equivalent technique to identify the most influential genes as PCA does with loadings. Our method, *InGene*, overcomes this hurdle and identifies genes that best explain non-linear transformations for single-cell transcriptome data [190]. *InGene* Gene-set enrichment analysis on top *InGene* genes for Melanoma and Human Breast Cancer shows that *InGene* genes have the best disease-gene association when compared with other supervised and unsupervised methods. This proves that *InGene* can provide reliable gene panel for targeted sequencing for scRNA-seq data thus reducing the cost manifold. Using a cost-effective scRNA-seq sequencing solution can prove to be a headway in personalized therapy recommendation and help make the clinical decision making process more effective. *InGene* not only provides an equivalent technique to “PCA loadings” for non-linear methods but also successfully extracts

biologically relevant genes for the tested datasets. *InGene* is fast and scalable. The associated gene ranking method is free of any parametric assumption, which makes *InGene* assay-independent.

5.1.2 Developing machine learning based strategies to design a non-invasive, blood-based inexpensive, cancer screening technology

In cancer, platelet transcriptome undergoes significant changes, providing a remarkable opportunity to utilize them in devising novel diagnostic strategies. Problems with these approaches are two fold. First, an optimal disease prediction often requires several tens of genes. Profiling large numbers of genes incur a significant cost, impeding its diagnostic adoption. We asked if the gene set could be narrowed down without compromising disease predictability. Secondly, the validation of a gene signature is contingent on the availability of a large number of tissue samples. To overcome these limitations, we developed a pipeline that maximizes disease-healthy classification performance with limited feature genes and a small validation cohort. We successfully augmented the validation cohort with artificial data points, significantly boosting the classification accuracy [191]. This could be useful in many practical scenarios where low sample acquisition rates impede the study's progress and clinical adoption. Some other conditions, like myocardial infarction (MI), are also known to cause changes in the platelet transcriptome. We tested our gene panel on myocardial infarction data, and the gene signature failed to classify MI patients from healthy ones, thus indicating a cancer-specific gene signature. Notably, the selection of these 11 genes was not biased to any particular cancer and, therefore, can be used across at least four other cancer types other than non-small-cell lung cancer (NSCLC). These include colorectal (CRC), glioblastoma multiforme (GBM), breast cancer (BRCA), and pancreatic cancer (PC). We addressed some critical real-world challenges in this study - experimental validation, affordability, reproducibility on geographically distinct populations, and disease specificity.

5.1.3 Developing and validating a risk stratification model for early relapse of multiple myeloma after ASCT using machine learning

A lot of success in the treatment can be attributed to Autologous Stem Cell Transplantation (ASCT). However, many patients across various financial and social backgrounds hesitate to undergo ASCT. The two main attributing reasons are - 1) ASCT is a somewhat expensive procedure. 2) It is an invasive procedure. Thus, when faced with uncertainty about the success of the procedure, a significant number of patients opt out. In a retrospective analysis, we noted that the existing staging schemes (ISS and DSS) would be of little use in this scenario due to their poor correlation with the graft outcome. This inspired us to explore the potential of multivariate modelling of the outcome of stem cell rescue in MM. We developed a new machine learning approach that's minimally susceptible to the problem of model overfitting. We showed that a simple, 3-variable (if relapse occurs after remission, response to induction and (pre-transplant) GFR) decision tree could serve as a staging scheme that maps each patient to one of the two (high and low) risk groups, with markedly distinct survival patterns for both overall and progression-free survival [192]. The benefit of our model is two-pronged:

- It is a simple, easy-to-understand staging system. Often clinicians would not find the time to run a complex, data-hungry, time taking model. We mitigate this issue by coming up with a model that is small and easy enough to memorize. It requires little to no computational power to bucket a patient in the "high risk" or "low risk" group. Our staging system reflects the power of interpretable models in real-world scenarios.
- When a clinician finds a patient hesitant to undergo ASCT, they could use the model to generate the likely outcome. If the model generates a favourable outcome should the patient undergo ASCT, they could discuss the future treatment path with the patient, aided by a data-driven hypothesis.

In conclusion, this thesis comprehensively explores innovative machine learning techniques in cancer diagnostics and treatment. From developing a non-invasive, blood-based cancer screening technology to identifying influential genes in complex datasets and, finally, devising a risk stratification model for early relapse of multiple myeloma,

each chapter contributes to the overarching goal of enhancing personalized medicine. These advancements not only demonstrate the potential of machine learning in tackling critical challenges in oncology but also pave the way for future research aimed at refining and validating these methods for clinical application.

5.2 Future Work

- We would like to perform whole transcriptome analysis on 500 NSCLC and 250 healthy samples and discern if there exists a gene panel specific to the Indian diaspora. This will allow us to focus on the Indian populace suffering from NSCLC and make strides towards precision medicine in the real sense.
- There is also potential in integrating TEP genes with the cancer gene panels. TEP signatures could complement existing gene panels by adding specific markers relevant to NSCLC, potentially enhancing diagnostic accuracy and early detection capabilities. However, integrating TEP signatures with NGS-based cancer panels would increase the complexity of both the assay and the data analysis. The TEP signatures would require extensive validation to be integrated into existing NGS panels, which are already standardized for clinical use.
- Spatial transcriptomics is a highly effective method that has significantly enhanced our comprehension of cellular interactions and tissue functional organization. *InGene* captures some spatial information using tSNE and UMAP layout. However, the low-dimensional representation might not always be representative of the true spatial coordinates of the cells. Therefore, integrating *InGene* gene signatures with spatial coordinates will aid in a more comprehensive understanding of single-cell subpopulation.

REFERENCES

[1] Janice S Dorman, Mandy J Schmella, and Susan W Wesmiller. Primer in genetics and genomics, article 1: Dna, genes, and chromosomes. Biological Research for Nursing, 19(1):7–17, 2017.

[2] NIH. How do genes direct the production of proteins? URL https://medlineplus.gov/genetics/understanding/howgeneswork/makingprotein/.

[3] Terence A Brown. Mutation, repair and recombination. In Genomes. 2nd edition. Wiley-Liss, 2002.

[4] Alex Vassilev and Melvin L DePamphilis. Links between dna replication, stem cells and cancer. Genes, 8(2):45, 2017.

[5] C Athena Aktipis and Randolph M Nesse. Evolutionary foundations for cancer biology. Evolutionary applications, 6(1):144–159, 2013.

[6] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. cell, 144(5):646–674, 2011.

[7] www.cancer.org. Understanding what cancer is: Ancient times to present. URL https://www.cancer.org/treatment/understanding-your-diagnosis/history-of-cancer/what-is-cancer.html.

[8] Arne Hessenbruch. A brief history of x-rays. Endeavour, 26(4):137–141, 2002.

[9] Vincent T DeVita Jr and Edward Chu. A history of cancer chemotherapy. Cancer research, 68(21):8643–8653, 2008.

[10] Shah Hussain, Iqra Mubeen, Niamat Ullah, Syed Shahab Ud Din Shah, Bakhtawar Abduljalil Khan, Muhammad Zahoor, Riaz Ullah, Farhat Ali Khan, and Mujeeb A Sultan. Modern diagnostic imaging technique applications and risk factors in the medical field: A review. BioMed Research International, 2022, 2022.

[11] R Scott Evans. Electronic health records: then, now, and in the future. Yearbook of medical informatics, 25(S 01):S48–S61, 2016.

[12] Christina L Clarke and Heather S Feigelson. Developing an algorithm to identify history of cancer using electronic medical records. eGEMs, 4(1), 2016.

[13] Veronique G LeBlanc and Marco A Marra. Next-generation sequencing approaches in cancer: where have they brought us and where will they take us? Cancers, 7(3):1925–1958, 2015.

[14] Saife N Lone, Sabah Nisar, Tariq Masoodi, Mayank Singh, Arshi Rizwan, Sheema Hashem, Wael El-Rifai, Davide Bedognetti, Surinder K Batra, Mohammad Haris, et al. Liquid biopsy: A step closer to transform diagnosis, prognosis and future of cancer treatments. Molecular cancer, 21(1):1–22, 2022.

[15] Roberta Pastorino, Corrado De Vito, Giuseppe Migliara, Katrin Glocker, Ilona Binenbaum, Walter Ricciardi, and Stefania Boccia. Benefits and challenges of big data in healthcare: an overview of the european initiatives. European journal of public health, 29(Supplement_3):23–27, 2019.

[16] Ying Xie, Wei-Yu Meng, Run-Ze Li, Yu-Wei Wang, Xin Qian, Chang Chan, Zhi-Fang Yu, Xing-Xing Fan, Hu-Dan Pan, Chun Xie, et al. Early lung cancer diagnostic biomarker discovery by machine learning methods. Translational oncology, 14(1):100907, 2021.

[17] Sandra Steyaert, Marija Pizurica, Divya Nagaraj, Priya Khandelwal, Tina Hernandez-Boussard, Andrew J Gentles, and Olivier Gevaert. Multimodal data fusion for cancer biomarker discovery with deep learning. Nature Machine Intelligence, pages 1–12, 2023.

[18] IOS Press. Ai and machine learning could improve cancer diagnosis through biomarker discovery. URL https://www.news-medical.net/news/20220301/AI-and-machine-learning-could-improve-cancer-diagnosis-through-aspx.

[19] András Lánczky and Balázs Győrffy. Web-based survival analysis tool tailored for medical research (kmplot): development and implementation. Journal of medical Internet research, 23(7):e27633, 2021.

[20] Laura Dean. Tamoxifen therapy and cyp2d6 genotype. 2019.

[21] Jina Ko, Steven N Baldassano, Po-Ling Loh, Konrad Kording, Brian Litt, and David Issadore. Machine learning to detect signatures of disease in liquid biopsies–a user's guide. Lab on a Chip, 18(3):395–405, 2018.

[22] Linjing Liu, Xingjian Chen, Olutomilayo Olayemi Petinrin, Weitong Zhang, Saifur Rahaman, Zhi-Ri Tang, and Ka-Chun Wong. Machine learning protocols in early cancer detection based on liquid biopsy: a survey. Life, 11(7):638, 2021.

[23] Tianyu Wang, Jun Bai, and Sheida Nabavi. Single-cell classification using graph convolutional networks. BMC bioinformatics, 22(1):1–23, 2021.

[24] Qin Ma and Dong Xu. Deep learning shapes single-cell data analysis. Nature Reviews Molecular Cell Biology, 23(5):303–304, 2022.

[25] canceratlas.cancer.org. The economic burden of cancer. URL https://canceratlas.cancer.org/taking-action/economic-burden.

[26] Joohyun Park and Kevin A Look. Health care expenditure burden of cancer care in the united states. INQUIRY: The Journal of Health Care Organization, Provision, and Financing, 56:0046958019880696, 2019.

[27] Pricivel M Carrera, Hagop M Kantarjian, and Victoria S Blinder. The financial burden and distress of patients with cancer: understanding and stepping-up action on the financial toxicity of cancer treatment. CA: a cancer journal for clinicians, 68(2):153–165, 2018.

[28] Whitney Zahnd and Sabha Ganai. Access to cancer care in rural populations: barriers and solutions. ASCO Daily, 2019.

[29] K Robin Yabroff, Jennifer Lund, Deanna Kepka, and Angela Mariotto. Economic burden of cancer in the united states: estimates, projections, and future research. Cancer Epidemiology and Prevention Biomarkers, 20(10):2006–2014, 2011.

[30] Joan L Warren, K Robin Yabroff, Angela Meekins, Marie Topor, Elizabeth B Lamont, and Martin L Brown. Evaluation of trends in the cost of initial cancer treatment. Journal of the National Cancer Institute, 100(12):888–897, 2008.

[31] Peter B Bach. Limits on medicare's ability to control rising spending on cancer drugs. The New England journal of medicine, 360(6):626–633, 2009.

[32] Nicholas McGranahan and Charles Swanton. Clonal heterogeneity and tumor evolution: past, present, and the future. Cell, 168(4):613–628, 2017.

[33] Jialing Zhang, Stephan Stanislaw Späth, Sadie L Marjani, Wengeng Zhang, and Xinghua Pan. Characterization of cancer genomic heterogeneity by next-generation sequencing advances precision medicine in cancer treatment. Precision clinical medicine, 1(1):29–48, 2018.

[34] Vinay Prasad, Tito Fojo, and Michael Brada. Precision oncology: origins, optimism, and potential. The Lancet Oncology, 17(2):e81–e86, 2016.

[35] Brian J Druker, Moshe Talpaz, Debra J Resta, Bin Peng, Elisabeth Buchdunger, John M Ford, Nicholas B Lydon, Hagop Kantarjian, Renaud Capdeville, Sayuri Ohno-Jones, et al. Efficacy and safety of a specific inhibitor of the bcr-abl tyrosine kinase in chronic myeloid leukemia. New England Journal of Medicine, 344(14):1031–1037, 2001.

[36] Dennis J Slamon, Gary M Clark, Steven G Wong, Wendy J Levin, Axel Ullrich, and William L McGuire. Human breast cancer: correlation of relapse and survival with amplification of the her-2/neu oncogene. science, 235(4785):177–182, 1987.

[37] Dennis J Slamon, Brian Leyland-Jones, Steven Shak, Hank Fuchs, Virginia Paton, Alex Bajamonde, Thomas Fleming, Wolfgang Eiermann, Janet Wolter, Mark Pegram, et al. Use of chemotherapy plus a monoclonal antibody against her2 for metastatic breast cancer that overexpresses her2. New England journal of medicine, 344(11):783–792, 2001.

[38] Caroline Robert, Jacob Schachter, Georgina V Long, Ana Arance, Jean Jacques Grob, Laurent Mortier, Adil Daud, Matteo S Carlino, Catriona McNeil, Michal Lotem, et al. Pembrolizumab versus ipilimumab in advanced melanoma. New England Journal of Medicine, 372(26):2521–2532, 2015.

[39] Sara Huston Katsanis and Nicholas Katsanis. Molecular genetic testing and the future of clinical genomics. Nature Reviews Genetics, 14(6):415–426, 2013.

[40] Kimberly R Kukurba and Stephen B Montgomery. Rna sequencing and analysis. Cold Spring Harbor Protocols, 2015(11):pdb–top084970, 2015.

[41] Francesca Finotello and Barbara Di Camillo. Measuring differential gene expression with rna-seq: challenges and strategies for data analysis. Briefings in functional genomics, 14(2):130–142, 2015.

[42] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. Nature reviews genetics, 10(1):57–63, 2009.

[43] Dragomirka Jovic, Xue Liang, Hua Zeng, Lin Lin, Fengping Xu, and Yonglun Luo. Single-cell rna sequencing technologies and applications: A brief overview. Clinical and Translational Medicine, 12(3):e694, 2022.

[44] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell, 161(5):1202–1214, 2015.

[45] James A Briggs, Caleb Weinreb, Daniel E Wagner, Sean Megason, Leonid Peshkin, Marc W Kirschner, and Allon M Klein. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. Science, 360(6392): eaar5780, 2018.

[46] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of single-cell rna sequencing methods. Molecular cell, 65(4):631–643, 2017.

[47] Blue B Lake, Rizi Ai, Gwendolyn E Kaeser, Neeraj S Salathia, Yun C Yung, Rui Liu, Andre Wildberg, Derek Gao, Ho-Lim Fung, Song Chen, et al. Neuronal subtypes and diversity revealed by single-nucleus rna sequencing of the human brain. Science, 352(6293):1586–1590, 2016.

[48] Rashel V Grindberg, Joyclyn L Yee-Greenbaum, Michael J McConnell, Mark Novotny, Andy L O'Shaughnessy, Georgina M Lambert, Marcos J Araúzo-Bravo, Jun Lee, Max Fishman, Gillian E Robbins, et al. Rna-sequencing from single nuclei. Proceedings of the National Academy of Sciences, 110(49): 19802–19807, 2013.

[49] Suguna Rani Krishnaswami, Rashel V Grindberg, Mark Novotny, Pratap Venepally, Benjamin Lacar, Kunal Bhutani, Sara B Linker, Son Pham, Jennifer A Erwin, Jeremy A Miller, et al. Using single nuclei for rna-seq to capture the transcriptome of postmortem neurons. Nature protocols, 11(3):499–524, 2016.

[50] John H Holmes, James Beinlich, Mary R Boland, Kathryn H Bowles, Yong Chen, Tessa S Cook, George Demiris, Michael Draugelis, Laura Fluharty, Peter E Gabriel, et al. Why is the electronic health record so challenging for research and clinical care? Methods of information in medicine, 60(01/02):032–048, 2021.

[51] Brett K Beaulieu-Jones, Jason H Moore, and POOLED RESOURCE OPEN-ACCESS ALS CLINICAL TRIALS CONSORTIUM. Missing data imputation in the electronic health record using deeply learned autoencoders. In Pacific symposium on biocomputing 2017, pages 207–218. World Scientific, 2017.

[52] Visar Berisha, Chelsea Krantsevich, P Richard Hahn, Shira Hahn, Gautam Dasarathy, Pavan Turaga, and Julie Liss. Digital medicine and the curse of dimensionality. NPJ digital medicine, 4(1):153, 2021.

[53] Jason Brownlee. Overfitting and underfitting with machine learning algorithms. Machine Learning Mastery, 21:575, 2016.

[54] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. Journal of machine learning research, 3(Mar):1157–1182, 2003.

[55] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. Genome biology, 21(1):1–35, 2020.

[56] Namrata Bhattacharya, Colleen C Nelson, Gaurav Ahuja, and Debarka Sengupta. Big data analytics in single-cell transcriptomics: Five grand opportunities. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(4):e1414, 2021.

[57] Ruochen Jiang, Tianyi Sun, Dongyuan Song, and Jingyi Jessica Li. Statistics or biology: the zero-inflation controversy about scrna-seq data. Genome biology, 23(1):1–24, 2022.

[58] Hung-I Harry Chen, Yufang Jin, Yufei Huang, and Yidong Chen. Detection of high variability in gene expression from single-cell rna-seq profiling. BMC genomics, 17:119–128, 2016.

[59] Jie Sheng and Wei Vivian Li. Selecting gene features for unsupervised analysis of single-cell gene expression data. Briefings in bioinformatics, 22(6):bbab295, 2021.

[60] Saket Choudhary and Rahul Satija. Comparison and evaluation of statistical error models for scrna-seq. Genome biology, 23(1):27, 2022.

[61] Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell rna-seq denoising using a deep count autoencoder. Nature communications, 10(1):390, 2019.

[62] Andreas Janecek, Wilfried Gansterer, Michael Demel, and Gerhard Ecker. On the relationship between feature selection and classification accuracy. In New Challenges for Feature Selection in Data Mining and Knowledge Discovery, pages 90–105, 2008.

[63] Yanxia Li, Yi Chai, Han Zhou, and Hongpeng Yin. A novel dimension reduction and dictionary learning framework for high-dimensional data classification. Pattern Recognition, 112:107793, 2021.

[64] Sean Simmons, Jian Peng, Jadwiga Bienkowska, and Bonnie Berger. Discovering what dimensionality reduction really tells us about rna-seq data. Journal of Computational Biology, 22(8):715–728, 2015.

[65] Heather J Zhou, Lei Li, Yumei Li, Wei Li, and Jingyi Jessica Li. Pca outperforms popular hidden variable inference methods for molecular qtl mapping. Genome Biology, 23(1):1–17, 2022.

[66] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008.

[67] Marco Russano, Andrea Napolitano, Giulia Ribelli, Michele Iuliani, Sonia Simonetti, Fabrizio Citarella, Francesco Pantano, Emanuela Dell'Aquila, Cecilia Anesi, Nicola Silvestris, et al. Liquid biopsy and tumor heterogeneity in metastatic solid tumors: the potentiality of blood samples. Journal of Experimental & Clinical Cancer Research, 39(1):1–13, 2020.

[68] Lawrence Hsu Lin, Douglas HR Allison, Yang Feng, George Jour, Kyung Park, Fang Zhou, Andre L Moreira, Guomiao Shen, Xiaojun Feng, Joshua Sabari, et al. Comparison of solid tissue sequencing and liquid biopsy accuracy in identification of clinically relevant gene mutations and rearrangements in lung adenocarcinomas. Modern Pathology, 34(12):2168–2174, 2021.

[69] Ernesto Lopez-Corona, Makoto Ohori, Thomas M Wheeler, Victor E Reuter, Peter T Scardino, Michael W Kattan, and James A Eastham. Prostate cancer diagnosed after repeat biopsies have a favorable pathological outcome but similar recurrence rate. The Journal of urology, 175(3):923–928, 2006.

[70] Jacob J Adashek, Filip Janku, and Razelle Kurzrock. Signed in blood: circulating tumor dna in cancer diagnosis, treatment and screening. Cancers, 13(14):3600, 2021.

[71] Douglas S Micalizzi, Shyamala Maheswaran, and Daniel A Haber. A conduit to metastasis: circulating tumor cell biology. Genes & development, 31(18): 1827–1840, 2017.

[72] Carlyn Rose C Tan, Lanlan Zhou, and Wafik S El-Deiry. Circulating tumor cells versus circulating tumor dna in colorectal cancer: pros and cons. Current colorectal cancer reports, 12:151–161, 2016.

[73] Ying Yang, Hongyang Liu, Youming Chen, Nan Xiao, Zhaoyang Zheng, Hongchun Liu, and Junhu Wan. Liquid biopsy on the horizon in immunotherapy of non-small cell lung cancer: current status, challenges, and perspectives. Cell Death & Disease, 14(3):230, 2023.

[74] Ruth Etzioni, Nicole Urban, Scott Ramsey, Martin McIntosh, Stephen Schwartz, Brian Reid, Jerald Radich, Garnet Anderson, and Leland Hartwell. The case for early detection. Nature reviews cancer, 3(4):243–252, 2003.

[75] Susan J Curry, Tim Byers, Maria Hewitt, et al. Potential of screening to reduce the burden of cancer. In Fulfilling the Potential of Cancer Prevention and Early Detection. National Academies Press (US), 2003.

[76] Sjors GJG, Thomas Wurdinger, et al. Tumor-educated platelets. Blood, 133(22): 2359–2364, 2019.

[77] Armand Trousseau. Lectures on clinical medicine, volume 2. Lindsay & Blakiston, 1873.

[78] T Billroth. Lectures on surgical pathology and therapeutics: A handbook for students and practitioners, ed 8 (translated). london. The Syndenham Society, 1878.

[79] Myron G Best, Nik Sol, Irsan Kooi, Jihane Tannous, Bart A Westerman, François Rustenburg, Pepijn Schellen, Heleen Verschueren, Edward Post, Jan Koster, Bauke Ylstra, Najim Ameziane, Josephine Dorsman, Egbert F Smit, Henk M Verheul, David P Noske, Jaap C Reijneveld, R Jonas A Nilsson, Bakhos A Tannous, Pieter Wesseling, and Thomas Wurdinger. RNA-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. Cancer Cell, 28(5):666–676, nov 2015. doi: 10.1016/j.ccell.2015.09.018. URL http://dx.doi.org/10.1016/j.ccell.2015.09.018.

[80] JACK LEVIN and C LOCKARD CONLEY. Thrombocytosis associated with malignant disease. Archives of internal medicine, 114(4):497–500, 1964.

[81] Martin Schlesinger. Role of platelets and platelet receptors in cancer metastasis. Journal of hematology & oncology, 11(1):1–15, 2018.

[82] Aashi Jindal, Prashant Gupta, Debarka Sengupta, et al. Discovery of rare cells from voluminous single cell expression data. Nature communications, 9(1):1–9, 2018.

[83] Marta Tellez-Gabriel, Benjamin Ory, Francois Lamoureux, Marie-Francoise Heymann, and Dominique Heymann. Tumour heterogeneity: the key advantages of single-cell analysis. International journal of molecular sciences, 17(12): 2142, 2016.

[84] Rhonda Bacher and Christina Kendziorski. Design and computational analysis of single-cell rna-sequencing experiments. Genome biology, 17(1):1–14, 2016.

[85] Martin Sewell. Principal component analysis. 2007.

[86] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(11), 2008.

[87] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426, 2018.

[88] Barbara Treutlein, Doug G Brownfield, Angela R Wu, Norma F Neff, Gary L Mantalas, F Hernan Espinoza, Tushar J Desai, Mark A Krasnow, and Stephen R Quake. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. Nature, 509(7500):371, 2014.

[89] Yan Wu, Pablo Tamayo, and Kun Zhang. Visualizing and interpreting single-cell gene expression datasets with similarity weighted nonnegative embedding. Cell systems, 7(6):656–666, 2018.

[90] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using umap. Nature biotechnology, 37(1):38–44, 2019.

[91] Spyros Darmanis, Steven A Sloan, Ye Zhang, Martin Enge, Christine Caneda, Lawrence M Shuer, Melanie G Hayden Gephart, Ben A Barres, and Stephen R Quake. A survey of human brain transcriptome diversity at the single cell level. Proceedings of the National Academy of Sciences, 112(23):7285–7290, 2015.

[92] Brian S Everitt and Anders Skrondal. The cambridge dictionary of statistics. 2010.

[93] Ugo Fano. Ionization yield of radiations. ii. the fluctuations of the number of ions. Physical Review, 72(1):26, 1947.

[94] Marina Wright Muelas, Farah Mughal, Steve O'Hagan, Philip J Day, and Douglas B Kell. The role and robustness of the gini coefficient as an unbiased tool for the selection of gini genes for normalising expression profiling data. Scientific reports, 9(1):1–21, 2019.

[95] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3):37–52, 1987.

[96] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. Genome biology, 16(1):1–13, 2015.

[97] Bianca Dumitrascu, Soledad Villar, Dustin G Mixon, and Barbara E Engelhardt. Optimal marker gene selection for cell type discrimination in single cell analyses. Nature communications, 12(1):1–8, 2021.

[98] Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. Science, 352(6282):189–196, 2016.

[99] 10x genomics. human breast cancer (block a section 1), 2019. https://support.10xgenomics.com/spatial-gene-expression/datasets/1.0.0/V1_Breast_Cancer_Block_A_Section_1.

[100] Fresh 68k pbmcs (donor a), 2016. https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/fresh_68k_pbmc_donor_a.

[101] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I Furlong. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic acids research, page gkw943, 2016.

[102] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. BMC bioinformatics, 14 (1):1–14, 2013.

[103] Parmanand Malvi, Radoslav Janostiak, Arvindhan Nagarajan, Xuchen Zhang, and Narendra Wajapeyee. N-acylsphingosine amidohydrolase 1 promotes melanoma growth and metastasis by suppressing peroxisome biogenesis-induced ros production. Molecular metabolism, 48:101217, 2021.

[104] Justine Leclerc, David Garandeau, Charlotte Pandiani, Céline Gaudel, Karine Bille, Nicolas Nottet, Virginie Garcia, Pascal Colosetti, Sophie Pagnotta, Philippe Bahadoran, et al. Lysosomal acid ceramidase asah1 controls the transition between invasive and proliferative phenotype in melanoma cells. Oncogene, 38(8):1282–1295, 2019.

[105] L Isabel Cárdenas-Navia, Pedro Cruz, Jimmy C Lin, NISC Comparative Sequencing Program, Steven A Rosenberg, Yardena Samuels, et al. Novel somatic mutations in heterotrimeric g proteins in melanoma. Cancer biology & therapy, 10(1):33–37, 2010.

[106] Isabelle R Cohen, Alan D Murdoch, Michael F Naso, Dario Marchetti, David Berd, and Renato V Iozzo. Abnormal expression of perlecan proteoglycan in metastatic melanomas. Cancer research, 54(22):5771–5774, 1994.

[107] Wenjing Zhang, Zhijuan Lin, Fuyan Shi, Qiang Wang, Yujia Kong, Yanfeng Ren, Juncheng Lyu, Chao Sheng, Yuting Li, Hao Qin, et al. Hspg2 mutation association with immune checkpoint inhibitor outcome in melanoma and non-small cell lung cancer. Cancers, 14(14):3495, 2022.

[108] Chunrong Song, Zhong Su, and Jing Guo. Thymosin β 10 is overexpressed and associated with unfavorable prognosis in hepatocellular carcinoma. Bioscience reports, 39(3), 2019.

[109] William M Hardesty, Mark C Kelley, Deming Mi, Robert L Low, and Richard M Caprioli. Protein signatures for survival and recurrence in metastatic melanoma. Journal of proteomics, 74(7):1002–1014, 2011.

[110] Marian AJ Weterman, Goos NP Van Muijen, Dirk J Ruiter, and Henri PJ Bloemers. Thymosin β-10 expression in melanoma cell lines and melanocytic lesions: A new progression marker for human cutaneous melanoma. International journal of cancer, 53(2):278–284, 1993.

[111] Lih Yin Tan, Chris Mintoff, M Zahied Johan, Brenton W Ebert, Clare Fedele, You Fang Zhang, Pacman Szeto, Karen E Sheppard, Grant A McArthur, Erwin Foster-Smith, et al. Desmoglein 2 promotes vasculogenic mimicry in melanoma and is associated with poor clinical outcome. Oncotarget, 7(29):46492, 2016.

[112] Wiebke K Peitsch, Yvette Doerflinger, Reiner Fischer-Colbrie, Volker Huck, Alexander T Bauer, Jochen Utikal, Sergij Goerdt, and Stefan W Schneider. Desmoglein 2 depletion leads to increased migration and upregulation of the chemoattractant secretoneurin in melanoma cells. PLoS One, 9(2):e89491, 2014.

[113] Valentine Svensson, Sarah A Teichmann, and Oliver Stegle. Spatialde: identification of spatially variable genes. Nature methods, 15(5):343–346, 2018.

[114] Jinhua Xu, Yinghua Chen, and Olufunmilayo I Olopade. Myc and breast cancer. Genes & cancer, 1(6):629–640, 2010.

[115] Yassi Fallah, Janetta Brundage, Paul Allegakoen, and Ayesha N Shajahan-Haq. Myc-driven pathways in breast cancer subtypes. Biomolecules, 7(3):53, 2017.

[116] Li Xu Yan, Qi Nian Wu, Yan Zhang, Yang Yang Li, Ding Zhun Liao, Jing Hui Hou, Jia Fu, Mu Sheng Zeng, Jing Ping Yun, Qiu Liang Wu, et al. Knockdown of mir-21 in human breast cancer cell lines inhibits proliferation, in vitro migration and in vivotumor growth. Breast cancer research, 13(1):1–14, 2011.

[117] Mei Liu, Can Gong, Renyuan Xu, Yu Chen, and Xiaodong Wang. Microrna-5195-3p enhances the chemosensitivity of triple-negative breast cancer to paclitaxel by downregulating eif4a2. Cellular & Molecular Biology Letters, 24(1): 1–11, 2019.

[118] Zehra Elgundi, Michael Papanicolaou, Gretel Major, Thomas R Cox, James Melrose, John M Whitelock, and Brooke L Farrugia. Cancer metastasis: the role of the extracellular matrix and the heparan sulfate proteoglycan perlecan. Frontiers in oncology, 9:1482, 2020.

[119] Renato V Iozzo and Ralph D Sanderson. Proteoglycans in cancer biology, tumour microenvironment and angiogenesis. Journal of cellular and molecular medicine, 15(5):1013–1031, 2011.

[120] Antonio J Giráldez, Richard R Copley, and Stephen M Cohen. Hspg modification by the secreted enzyme notum shapes the wingless morphogen gradient. Developmental cell, 2(5):667–676, 2002.

[121] Stephen Kalscheuer, Vidhi Khanna, Hyunjoon Kim, Sihan Li, Deepali Sachdev, Arthur DeCarlo, Da Yang, and Jayanth Panyam. Discovery of hspg2 (perlecan) as a therapeutic target in triple negative breast cancer. Scientific reports, 9(1): 1–11, 2019.

[122] Vidhi Khanna, Stephen Kalscheuer, Jayanth Panyam, Da Yang, and Sihan Li. Therapeutic efficacy of antibodies targeting domain 1 of hspg2 in breast cancer. Cancer Research, 78(13_Supplement):2897–2897, 2018.

[123] Chen Zhang, Bingfei Xu, Shi Lu, Ying Zhao, and Pian Liu. Hn1 contributes to migration, invasion, and tumorigenesis of breast cancer by enhancing myc activity. Molecular cancer, 16(1):1–10, 2017.

[124] Yi Liu, Dong Soon Choi, Jianting Sheng, Joe E Ensor, Diana Hwang Liang, Cristian Rodriguez-Aguayo, Amanda Polley, Steve Benz, Olivier Elemento, Akanksha Verma, et al. Hn1l promotes triple-negative breast cancer stem cells through lepr-stat3 pathway. Stem cell reports, 10(1):212–227, 2018.

[125] Carolina Di Benedetto, Justin Oh, Zainab Choudhery, Weiquan Shi, Gilmer Valdes, and Paola Betancur. Nsmce2, a novel super-enhancer regulated gene, is linked to poor prognosis and therapy resistance in breast cancer. bioRxiv, 2022.

[126] Vincent A Traag, Ludo Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. Scientific reports, 9(1):1–12, 2019.

[127] Luca Scrucca, Michael Fop, T Brendan Murphy, and Adrian E Raftery. mclust 5: clustering, classification and density estimation using gaussian finite mixture models. The R journal, 8(1):289, 2016.

[128] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. IEEE Intelligent Systems and their applications, 13(4):18–28, 1998.

[129] Leo Breiman. Random forests. Springer Science and Business Media LLC, 2001. doi: 10.1023/a:1010933404324. URL http://link.springer.com/10.1023/A:1010933404324.

[130] Marvin N Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. arXiv preprint arXiv:1508.04409, 2015.

[131] Fuminori Sakurai, Nobuhiro Narii, Kyoko Tomita, Shinsaku Togo, Kazuhisa Takahashi, Mitsuhiro Machitani, Masashi Tachibana, Masaaki Ouchi, Nobuyoshi Katagiri, Yasuo Urata, et al. Efficient detection of human circulating tumor cells without significant production of false-positive cells by a novel conditionally replicating adenovirus. Molecular Therapy-Methods & Clinical Development, 3:16001, 2016.

[132] Suzanne Jenkins, James CH Yang, Suresh S Ramalingam, Karen Yu, Sabina Patel, Susie Weston, Rachel Hodge, Mireille Cantarini, Pasi A Jänne, Tetsuya Mitsudomi, et al. Plasma ctdna analysis for detection of the egfr t790m mutation in patients with advanced non–small cell lung cancer. Journal of Thoracic Oncology, 12(7):1061–1070, 2017.

[133] Simon A Joosse and Klaus Pantel. Tumor-educated platelets as liquid biopsy in cancer patients. Cancer Cell, 28(5):552–554, nov 2015. doi: 10.1016/j.ccell.2015.10.007. URL http://dx.doi.org/10.1016/j.ccell.2015.10.007.

[134] Myron G Best, Nik Sol, Sjors G J G In 't Veld, Adrienne Vancura, Mirte Muller, Anna-Larissa N Niemeijer, Aniko V Fejes, Lee-Ann Tjon Kon Fat, Anna E Huis In 't Veld, Cyra Leurs, Tessa Y Le Large, Laura L Meijer, Irsan E Kooi, François Rustenburg, Pepijn Schellen, Heleen Verschueren, Edward Post, Laurine E Wedekind, Jillian Bracht, Michelle Esenkbrink, Leon Wils, Francesca Favaro, Jilian D Schoonhoven, Jihane Tannous, Hanne Meijers-Heijboer, Geert Kazemier, Elisa Giovannetti, Jaap C Reijneveld, Sander Idema, Joep Killestein, Michal Heger, Saskia C de Jager, Rolf T Urbanus, Imo E Hoefer, Gerard Pasterkamp, Christine Mannhalter, Jose Gomez-Arroyo, Harm-Jan Bogaard, David P Noske, W Peter Vandertop, Daan van den Broek, Bauke Ylstra, R Jonas A Nilsson, Pieter Wesseling, Niki Karachaliou, Rafael Rosell, Elizabeth Lee-Lewandrowski, Kent B Lewandrowski, Bakhos A Tannous, Adrianus J de Langen, Egbert F Smit, Michel M van den Heuvel, and Thomas Wurdinger. Swarm intelligence-enhanced detection of non-small-cell lung cancer using tumor-educated platelets. Cancer Cell, 32(2):238–252.e9, aug 2017. doi: 10.1016/j.ccell.2017.07.004. URL http://dx.doi.org/10.1016/j.ccell.2017.07.004.

[135] Meiling Sheng, Zhaohui Dong, and Yanping Xie. Identification of tumor-educated platelet biomarkers of non-small-cell lung cancer. OncoTargets and therapy, 11:8143–8151, nov 2018. doi: 10.2147/{OTT}.S177384. URL http://dx.doi.org/10.2147/{OTT}.S177384.

[136] Myron G Best, Nik Sol, Sjors GJG, Adrienne Vancura, Mirte Muller, Anna-Larissa N Niemeijer, Aniko V Fejes, Lee-Ann Tjon Kon Fat, Anna E Huis, Cyra Leurs, et al. Swarm intelligence-enhanced detection of non-small-cell lung cancer using tumor-educated platelets. Cancer cell, 32(2):238–252, 2017.

[137] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1):139–140, jan 2010. doi: 10.1093/bioinformatics/btp616. URL http://dx.doi.org/10.1093/bioinformatics/btp616.

[138] Giuliana Gobbi, Cecilia Carubbi, Guidantonio Malagoli Tagliazucchi, Elena Masselli, Prisco Mirandola, Filippo Pigazzani, Antonio Crocamo, Maria Francesca Notarangelo, Sergio Suma, Elvezia Paraboschi, Giuseppe Maglietta, Srikanth Nagalla, Giulia Pozzi, Daniela Galli, Mauro Vaccarezza, Paolo Fortina, Sankar Addya, Adam Ertel, Paul Bray, Stefano Duga, Carlo Berzuini, Marco Vitale, and Diego Ardissino. Sighting acute myocardial infarction through platelet gene expression. Scientific Reports, 9(1):19574, dec 2019. doi: 10.1038/s41598-019-56047-0. URL http://dx.doi.org/10.1038/s41598-019-56047-0.

[139] Jerome H. Friedman. Greedy function approximation: a gradient boosting machine. The Annals of Statistics, 29(5):1189–1232, oct 2001. ISSN 0090-5364. doi: 10.1214/aos/1013203451. URL http://projecteuclid.org/euclid.aos/1013203451.

[140] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.R. Mullers. Fisher discriminant analysis with kernels. In Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468), pages 41–48. IEEE, 1999. ISBN 0-7803-5673-X. doi: 10.1109/{NNSP}.1999.788121. URL http://ieeexplore.ieee.org/document/788121/.

[141] Thomas D Schmittgen and Kenneth J Livak. Analyzing real-time PCR data by the comparative c(t) method. Nature Protocols, 3(6):1101–1108, 2008. ISSN 1750-2799. doi: 10.1038/nprot.2008.73. URL http://dx.doi.org/10.1038/nprot.2008.73.

[142] Chao-Chung Kuo, Sonja Hänzelmann, Nevcin Sentürk Cetin, Stefan Frank, Barna Zajzon, Jens-Peter Derks, Vijay Suresh Akhade, Gaurav Ahuja, Chandrasekhar Kanduri, Ingrid Grummt, et al. Detection of rna–dna binding sites in long noncoding rnas. Nucleic acids research, 47(6):e32–e32, 2019.

[143] Stefan Frank, Gaurav Ahuja, Deniz Bartsch, Nicole Russ, Wenjie Yao, Joseph Chao-Chung Kuo, Jens-Peter Derks, Vijay Suresh Akhade, Yulia Kargapolova, Theodore Georgomanolis, et al. yylnct defines a class of divergently transcribed lncrnas and safeguards the t-mediated mesodermal commitment of human pscs. Cell Stem Cell, 24(2):318–327, 2019.

[144] Jayadeva, Sumit Soman, and Soumya Saxena. Eigensample: A non-iterative technique for adding samples to small datasets. Appl. Soft Comput., 70:1064–1077, 2017.

[145] Agnieszka Mikołajczyk and Michał Grochowski. Data augmentation for improving deep learning in image classification problem. In 2018 international interdisciplinary PhD workshop (IIPhDW), pages 117–122. IEEE, 2018.

[146] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340, 2017.

[147] René Dreos, Giovanna Ambrosini, Romain Groux, Rouaïda Cavin Périer, and Philipp Bucher. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. Nucleic Acids Research, 45(D1):D51–D55, jan 2017. doi: 10.1093/nar/gkw1069. URL http://dx.doi.org/10.1093/nar/gkw1069.

[148] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research, 37(Web Server issue):W202–8, jul 2009. doi: 10.1093/nar/gkp335. URL http://dx.doi.org/10.1093/nar/gkp335.

[149] Aziz Khan, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A Castro-Mondragon, Robin van der Lee, Adrien Bessy, Jeanne Chèneby, Shubhada R Kulkarni, Ge Tan, Damir Baranasic, David J Arenillas, Albin Sandelin, Klaas Vandepoele, Boris Lenhard, Benoît Ballester, Wyeth W Wasserman, François Parcy, and Anthony Mathelier. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Research, 46(D1):D260–D266, jan 2018. doi: 10.1093/nar/gkx1126. URL http://dx.doi.org/10.1093/nar/gkx1126.

[150] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1):139–140, 2010.

[151] Robert A. Kyle and S. Vincent Rajkumar. Multiple myeloma. New England Journal of Medicine, 351(18):1860–1873, 2004. doi: 10.1056/NEJMra041875. URL https://doi.org/10.1056/NEJMra041875. PMID: 15509819.

[152] Antonio Palumbo and Kenneth Anderson. Multiple myeloma. New England Journal of Medicine, 364(11):1046–1060, 2011. doi: 10.1056/NEJMra1011442. URL https://doi.org/10.1056/NEJMra1011442. PMID: 21410373.

[153] Dickran Kazandjian. Multiple myeloma epidemiology and survival: A unique malignancy. In Seminars in oncology, volume 43, pages 676–681. Elsevier, 2016.

[154] Dean Smith and Kwee Yong. Multiple myeloma. BMJ, 346, 2013. doi: 10.1136/bmj.f3863. URL https://www.bmj.com/content/346/bmj.f3863.

[155] Christina Fitzmaurice, Christine Allen, Ryan M Barber, Lars Barregard, Zulfiqar A Bhutta, Hermann Brenner, Daniel J Dicker, Odgerel Chimed-Orchir, Rakhi Dandona, Lalit Dandona, et al. 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, 2017.

[156] D Max Parkin, Freddie Bray, J Ferlay, and Paola Pisani. Global cancer statistics, 2002. CA: a cancer journal for clinicians, 55(2):74–108, 2005.

[157] Adam J Waxman, Pamela J Mink, Susan S Devesa, William F Anderson, Brendan M Weiss, Sigurdur Y Kristinsson, Katherine A McGlynn, and Ola Landgren. Racial disparities in incidence and outcome in multiple myeloma: a population-based study. Blood, pages blood–2010, 2010.

[158] Ingemar Turesson, Ramon Velez, Sigurdur Y Kristinsson, and Ola Landgren. Patterns of improved survival in patients with multiple myeloma in the twenty-first century: a population-based study. Journal of Clinical Oncology, 28(5):830, 2010.

[159] Shaji K Kumar, S Vincent Rajkumar, Angela Dispenzieri, Martha Q Lacy, Suzanne R Hayman, Francis K Buadi, Steven R Zeldenrust, David Dingli, Stephen J Russell, John A Lust, et al. Improved survival in multiple myeloma and the impact of novel therapies. Blood, 111(5):2516–2520, 2008.

[160] Francesca Gay, Stefania Oliva, Maria Teresa Petrucci, Concetta Conticello, Lucio Catalano, Paolo Corradini, Agostina Siniscalchi, Valeria Magarotto, Luděk Pour, Angelo Carella, et al. 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. The lancet oncology, 16(16):1617–1629, 2015.

[161] Nicola Lehners, Natalia Becker, Axel Benner, Maria Pritsch, Martin Löpprich, Elias Karl Mai, Jens Hillengass, Hartmut Goldschmidt, and Marc-Steffen Raab. Analysis of long-term survival in multiple myeloma after first-line autologous stem cell transplantation: impact of clinical risk factors and sustained response. Cancer medicine, 7(2):307–316, 2018.

[162] J Anthony Child, Gareth J Morgan, Faith E Davies, Roger G Owen, Susan E Bell, Kim Hawkins, Julia Brown, Mark T Drayson, and Peter J Selby. High-dose chemotherapy with hematopoietic stem-cell rescue for multiple myeloma. New England Journal of Medicine, 348(19):1875–1883, 2003.

[163] Roberto Mina, Alessandra Larocca, Massimo Offidani, Sara Bringhen, Tommaso Caravita, Valeria Magarotto, Lucia Pantani, Francesco Di Raimondo, Alberto Bosi, Iolanda D Vincelli, et al. 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, 2015.

[164] A Gupta, L Kumar, D Dabkara, D Gupta, O Sharma, and V Sreeniwas. Multiple myeloma: Autologous stem cell transplantation versus conventional chemotherapy—a retrospective age and stage matched analysis. Journal of Clinical Oncology, 27(15_suppl):7041–7041, 2009.

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

[166] Scott I Bearman, FR Appelbaum, CD Buckner, FB Petersen, LD Fisher, RA Clift, and ED Thomas. Regimen-related toxicity in patients undergoing bone marrow transplantation. Journal of Clinical Oncology, 6(10):1562–1568, 1988.

[167] Brian GM Durie and Sydney E Salmon. 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, 1975.

[168] Philip R Greipp, Jesus San Miguel, Brian GM Durie, John J Crowley, Bart Barlogie, Joan Bladé, Mario Boccadoro, J Anthony Child, Hervé Avet-Loiseau, Robert A Kyle, et al. International staging system for multiple myeloma. Journal of clinical oncology, 23(15):3412–3420, 2005.

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

[170] Lalit Kumar, Rakesh Reddy Boya, Rohit Pai, P Harish, Anjali Mookerjee, B Sainath, Mukesh Bhimrao Patekar, Ranjit Kumar Sahoo, Prabhat Singh Malik, OD Sharma, et al. Autologous stem cell transplantation for multiple myeloma: Long-term results. The National medical journal of India, 29(4):192, 2016.

[171] Andrew S Levey, Juan P Bosch, Julia Breyer Lewis, Tom Greene, Nancy Rogers, and David Roth. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Annals of internal medicine, 130(6):461–470, 1999.

[172] Guido Ghilardi, Thomas Pabst, Barbara Jeker, Rouven Müller, Anne Cairoli, Antonia MS Müller, Mario Bargetzi, Felicitas Hitz, Helen Baldomero, Dominik Heim, et al. Melphalan dose in myeloma patients >=65 years of age undergoing high-dose therapy and autologous stem cell transplantation: a multicentric observational registry study. Bone marrow transplantation, page 1, 2018.

[173] S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. Journal of statistical software, pages 1–68, 2010.

[174] Meletios A Dimopoulos, Evangelos Terpos, Asher Chanan-Khan, Nelson Leung, Heinz Ludwig, Sundar Jagannath, Ruben Niesvizky, Sergio Giralt, Jean-Paul Fermand, Joan Bladé, et al. Renal impairment in patients with multiple myeloma: a consensus statement on behalf of the international myeloma working group. Journal of Clinical Oncology, 28(33):4976–4984, 2010.

[175] Tiina Podymow, Ahsan Alam, Murray Vasilevsky, Roch Beauchemin, Chaim Shustik, and Michael Sebag. A review of multiple myeloma patients on dialysis treated with high cutoff hemodialysis, 2010.

[176] Manish Kumar Goel, Pardeep Khanna, and Jugal Kishore. Understanding survival analysis: Kaplan-meier estimate. International journal of Ayurveda research, 1(4):274–278, 2010.

[177] Haizhou Wang and Mingzhou Song. Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming. The R journal, 3(2):29, 2011.

[178] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In Advances in neural information processing systems, pages 849–856, 2002.

[179] Peter M Todd and Gerd Gigerenzer. Précis of simple heuristics that make us smart. Behavioral and brain sciences, 23(5):727–741, 2000.

[180] Nathaniel D Phillips, Hansjörg Neth, Jan K Woike, and Wolfgang Gaissmaier. Fftrees: A toolbox to create, visualize, and evaluate fast-and-frugal decision trees. Judgment and Decision Making, 12(4):344–368, 2017.

[181] Imad Bou-Hamad, Denis Larocque, Hatem Ben-Ameur, et al. A review of survival trees. Statistics Surveys, 5:44–71, 2011.

[182] RF Cornell and AA Kassim. Evolving paradigms in the treatment of relapsed/refractory multiple myeloma: increased options and increased complexity. Bone marrow transplantation, 51(4):479, 2016.

[183] Sara Farshchi Zarabi, Esther Masih-Khan, Christine Chen, Vishal Kukreti, Anca Prica, Rodger Tiedemann, Suzanne Trudel, and Donna Reece. Results of salvage autologous stem cell transplantation (asct) for relapsed multiple myeloma (mm) in the era of novel agents: Outcome of patients (pts) receiving prior bortezomib (btz)-based therapy, 2016.

[184] Jean-Luc Harousseau, Herve Avet-Loiseau, Michel Attal, Catherine Charbonnel, Frederic Garban, Cyrille Hulin, Mauricette Michallet, Thierry Facon, Laurent Garderet, Gerald Marit, et al. Achievement of at least very good partial response is a simple and robust prognostic factor in patients with multiple myeloma treated with high-dose therapy: long-term analysis of the ifm 99-02 and 99-04 trials. Journal of Clinical Oncology, 27(34):5720–5726, 2009.

[185] Eliot C Heher, Helmut G Rennke, Jacob P Laubach, and Paul G Richardson. Kidney disease and multiple myeloma. Clinical Journal of the American Society of Nephrology, 8(11):2007–2017, 2013.

[186] Andrew S Levey, Josef Coresh, Ethan Balk, Annamaria T Kausz, Adeera Levin, Michael W Steffes, Ronald J Hogg, Ronald D Perrone, Joseph Lau, and Garabed Eknoyan. National kidney foundation practice guidelines for chronic kidney disease: evaluation, classification, and stratification. Annals of internal medicine, 139(2):137–147, 2003.

[187] Rama Al Hamed, Abdul Hamid Bazarbachi, Florent Malard, Jean-Luc Harousseau, and Mohamad Mohty. Current status of autologous stem cell transplantation for multiple myeloma. Blood cancer journal, 9(4):44, 2019.

[188] Justin LaPorte, Stacey Brown, Xu Zhang, Asad Bashey, Lawrence E Morris, H Kent Holland, and Scott R Solomon. Age related outcomes for multiple myeloma patients following autologous transplantation, 2014.

[189] Antonio Palumbo, Hervé Avet-Loiseau, Stefania Oliva, Henk M Lokhorst, Hartmut Goldschmidt, Laura Rosinol, Paul Richardson, Simona Caltagirone, Juan José Lahuerta, Thierry Facon, et al. Revised international staging system for multiple myeloma: a report from international myeloma working group. Journal of clinical oncology, 33(26):2863, 2015.

[190] Chitrita Goswami and Debarka Sengupta. Ingene: Finding influential genes from embeddings of nonlinear dimension reduction techniques. bioRxiv, pages 2023–06, 2023.

[191] Chitrita Goswami, Smriti Chawla, Deepshi Thakral, Himanshu Pant, Pramod Verma, Prabhat Singh Malik, Jayadeva, Ritu Gupta, Gaurav Ahuja, and Debarka Sengupta. Molecular signature comprising 11 platelet-genes enables accurate blood-based diagnosis of nsclc. BMC genomics, 21:1–12, 2020.

[192] Chitrita Goswami, Sarita Poonia, Lalit Kumar, and Debarka Sengupta. Staging system to predict the risk of relapse in multiple myeloma patients undergoing autologous stem cell transplantation. Frontiers in oncology, 9:633, 2019.