Analyzing genomic and clinical data of hematological malignancies by various computational techniques

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

AKANKSHA FARSWAN
PhD16106

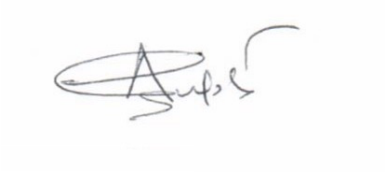
ADVISOR
PROF. ANUBHA GUPTA, IIIT DELHI

ELECTRONICS AND COMMUNICATION ENGINEERING
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI
NEW DELHI- 110020

February 14, 2023

Certificate

This is to certify that the thesis titled **Analyzing genomic and clinical data of hematological malignancies by various computational techniques**, submitted by **Akanksha Farswan**, to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of **Doctor of Philosophy**, is a bonafide record of the research work done by her under my 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.



Professor Anubha Gupta
Thesis Supervisor
Professor
SBILab, Dept. of Electronics and Communication
Indraprastha Institute of Information Technology, Delhi (IIITD)

Declaration

This is to certify that the thesis titled **Analyzing genomic and clinical data of hematological malignancies by various computational techniques**, submitted by me, to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of **Doctor of Philosophy**, is a bonafide work carried out by me. This research work has been carried out under the supervision of **Prof. Anubha Gupta**. The study pertaining to this thesis has not been submitted in part or in full, to any other University or Institution for the award of any other degree.

Akanksha Farswan
PhD Candidate
SBILab, Dept. of Electronics and Communication
Indraprastha Institute of Information Technology, Delhi (IIITD)

Acknowledgements

With the countless blessings of the Almighty God, it gives me immense joy to write that I have finally been able to accomplish my thesis. My Ph.D. journey has been the biggest learning curve of my life and it would not have been possible without the support of many people. I want to thank all of them for their help, knowing that words will fall short.

I would like to convey my profound gratitude to my advisor, Prof. Anubha Gupta, for believing in me and trusting me with the research work. She has been a constant source of inspiration, encouragement and support. Her enthusiasm for research has not only motivated me but also got me through the challenging times during my Ph.D. Her insightful feedback has transformed me into an independent and confident researcher. I shall be forever obliged for her kindness, valuable guidance and persistent help. I express my sincere gratitude to Prof. Ritu Gupta for her inventive inputs and recommendations, which immensely improved my research. She also provided essential resources for my work and connected me with excellent researchers at AIIMS. I extend my warmest thanks to Dr. Gurvinder Kaur and Dr. Lingaraja Jena for their insightful discussions, consistent help, and contributions throughout my Ph.D. journey.

I would also like to acknowledge Dr. Sriram K. for his precious time and for providing effective feedback on my work. I have learned a lot from the discussions with him. I sincerely thank my Ph.D. monitoring committee members, Prof. G. P. S. Raghava and Dr. Ganesh Bagler, for their fruitful observations and suggestions on my research work.

I express my heartfelt appreciation to my parents, Mrs. Suman Rani Farswan and Prof. Yogambar Singh Farswan, my sister, Jagriti Farswan, and my husband, Ankit Kaneri for their unconditional love and motivation. I am grateful to other family members, especially my grandfather-in-law, for their never-ending support, encouragement, and prayers that helped me sustain my research pursuit. Finally, I thank my friends Ashutosh and Shiv for all the helpful discussions and motivation during tough times of Ph.D.

Akanksha Farswan

Abstract

Large-scale characterization of the human genome has enabled the extensive study of the diverse genomic alterations present in humans. The integrative analyses of the various alterations provide a detailed understanding of the factors responsible for disease initiation and its progression in disorders like cancer. There is a wide range of machine learning algorithms and statistical methods to analyze genomic data and extract information for applications such as disease diagnosis and classification of clinical subtypes. These analyses assist in developing effective drugs for specific diseases and are particularly helpful in personalized cancer therapy, where the response of a patient to a particular drug can be captured, and its correlation with the mutation profiles of the patient can be examined to design targeted medicine. Though a plethora of methods exist for analyzing cancer genomes, certain challenges exist. Therefore, in this thesis, we have formulated and proposed different computational solutions to address challenges in cancer genomics, particularly in hematological malignancies.

Missing value problem is frequently observed in gene expression data, and it may significantly impact the findings extracted from the incomplete data. Therefore, we have dealt with the missing value in gene expression data by devising a compressive sensing (CS) based method, DSNN (Doubly Sparse in the Discrete Cosine Transform with Nuclear Norm minimization). A significant contribution is the utilization of Discrete Cosine Transform (DCT) based sparsity for recovering missing values. Further, we have analyzed the bulk-sequencing exome data of Multiple Myeloma (MM) and Monoclonal Gammopathy of Undetermined Significance (MGUS) patients. MM is a hematological cancer that arises from malignant transformation and deregulated proliferation of clonal plasma cells (PCs) in bone marrow, preceded by a benign condition of MGUS. The study has revealed actionable target genes that may be clinically relevant in addition to the genomic landscape of clonal evolution in MM. A statistically significant change in the mutational spectrum of MGUS and MM is observed as the disease progresses from MGUS and MM. We have also utilized survival data of the MM patients to find the association of Tumor mutational burden (TMB) with overall survival. In MM, it is critical to identify the initial risk stage of the patient as it helps in deciding the due course of the treatment to be given to the patient. Therefore, a reliable risk staging system is required, which may stratify the patients into separate subgroups and help identify patients requiring frequent visits to the hospital. Multiple staging systems have been proposed for MM, ISS and R-ISS being the gold standards used widely for

MM. However, none of them uses ethnicity information. Therefore, we have developed an ethnicity-aware Artificial Intelligence (AI)-enabled risk staging system, CRSS (Consensus-based Risk Staging System), for newly diagnosed multiple myeloma patients. The proposed method can predict the risk stage of any MM patient depending on the values of the simple parameters like age, albumin, β2-microglobulin, calcium, eGFR, hemoglobin and high-risk cytogenetic information. There has been an enhanced inclination towards single-cell sequencing data over bulk-sequencing data, given the several advantages of single-cell data over bulk NGS data. However, there are different noises present in the single-cell data. Therefore, in this thesis, we have devised an optimization-based framework, ARCANE-ROG, for denoising and imputing noisy and incomplete single-cell data for inferring patterns of clonal evolution.

Table of contents

List of tables

List of figures

Chapter 1

Introduction

1.1 Background

Cancer is a malignancy demonstrating an unrestricted proliferation of abnormal cells in the body. According to GLOBOCAN cancer statistics [1], 2020, there were an estimated 19.3 million new cases of cancer in the world in 2020. 10.0 million cancer mortalities were reported in the same year, making cancer one of the leading causes of death worldwide. Therefore, it is crucial to identify the mechanisms of cancer initiation, progression and relapse to improve the life expectancy among cancer patients. Cancers evolve and propagate via the acquisition of genetic mutations such as single nucleotide variants, small insertions/deletions and complex chromosomal aberrations like copy number variants and structural variants [2]. Advances in sequencing technology coupled with breakthroughs in computational approaches to store and analyze genomic data have enabled the large scale characterization of the human genome. After the completion of the first human genome project (HGP), numerous sequencing projects were initiated, such as the Human Genome Project–Write (HGP-Write) [3], which is a ten-year extension of the HGP, 100000 Genomes Project [4] and GenomeAsia 100K (GA100K) [5]. These projects are aimed to continue research on the human genome and unravel the genetic mysteries of diseases. Further, after the first cancer genome was sequenced [2], NGS data analysis has resulted in the creation of databases containing information on the mutations driving cancer or mutations which may be of potential clinical relevance. These databases include tumor alterations relevant for genomics-driven therapy (TARGET), Catalogue of Somatic Mutations in Cancer (COSMIC), and International Cancer Genomics Consortium (ICGC; https://dcc.icgc.org)

Conventional sequencing methods such as Sanger sequencing were expensive and time-consuming. However, next-generation sequencing (NGS) technologies, such as whole-exome sequencing (WES), whole-genome sequencing (WGS), RNA-sequencing (RNA-seq), etc., are high-throughput methods that support massively parallel sequencing of various genomic regions in multiple samples in a single run. In WGS, the entire genome is sequenced via a large DNA sample. Sequencing coverage for WGS should be high to detect clinically relevant mutations, which becomes expensive and time-consuming. On the other hand, WES focuses only on the coding regions (exons) of a genome which is nearly 2.5% of the entire human genome. Thus, WES is less costly and time-efficient than WGS, and it is more popularly used in cancer genomics for detecting rare and

common variants. RNA sequencing (RNA-Seq) assists in detecting changes in the gene expression profiles, alternative gene-spliced transcripts, gene fusion, etc. In addition to this, NGS technology is also used to investigate epigenetic alterations. Illumina/Solexa, SOLiD (Sequencing by Oligonucleotide Ligation and Detection) and Ion torrent are a few NGS platforms available today. PacBio sequencing and nanopore sequencing are referred to as third and fourth-generation sequencers. Though these techniques lag behind Illumina technology in terms of accuracy, they provide advantages over NGS, like longer read lengths. In addition, the 10x genomics technology introduced in 2016 enables the cell-by-cell analysis of the genome/transcriptome by using a Chromium system. Thus, the sequencing platforms are getting faster, more productive and cost-effective with time. Hence, a tremendous amount of NGS data is generated, demanding computational/bioinformatics skills to analyze the massive genomic data. Accordingly, there has been considerable development in the computational capacities to store and manage the data and computational methods to process the data.

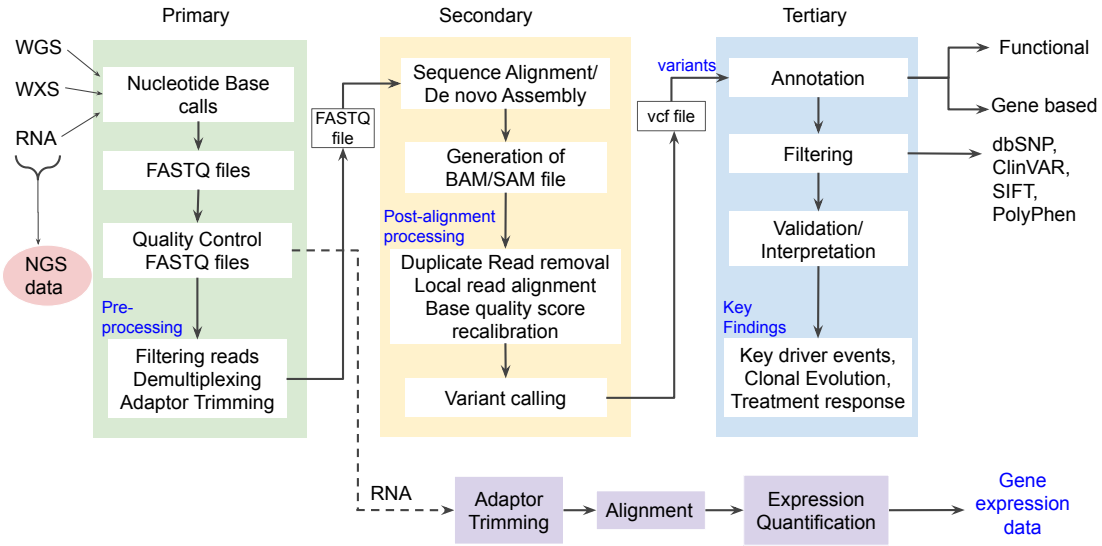


Figure 1.1: An overview of the NGS pipeline. NGS analysis is divided into three sections-Primary, Secondary and Tertiary analysis. The primary analysis mainly involves extracting nucleotide base calls from the raw data and converting them to FASTQ files. The quality check of FASTQ files is done to ensure high-quality reads, followed by pre-processing of FASTQ files. In the secondary analysis, pre-processed FASTQ files are aligned using a reference human genome or de novo assembly is done without a reference genome. BAM/SAM files are generated, which are further processed. Finally, variants are called. In the tertiary analysis, the variants are annotated and then filtered, and finally, the variants are interpreted using additional data to reveal novel findings from the data. This pipeline is mostly followed for exome or whole-genome files with few changes. For RNA-seq data, instead of calling variants, the expression quantification step is done to extract the gene expression profiles, which are analysed to get up-regulated or down-regulated genes.

All the steps involved in NGS data analysis can be categorized into three sections- primary, secondary, and tertiary, as shown in Figure 1.1. The primary analysis involves detecting the raw signal data and converting it into sequence data consisting of nucleotide base calls. Typically, Binary Base Call (BCL) files are the raw files generated from the sequencers, and these are converted to FASTQ files containing both the sequence data and the quality scores of the base call. Phred score, which represents the quality scores, is a logarithmic error probability. A Phred score of 10 (Q10) denotes an accuracy of 90%, i.e., a 1 in 10 probability of the base being incorrect. Similarly, Q30 means a 1 in 1000 probability of an incorrect base or 99.9% accuracy [6]. Thus, higher scores indicate high confidence in the base calls or better quality reads [7]. After generating FASTQ files, pre-processing of NGS reads is done to ensure that only high-quality reads are used for downstream analysis. Pre-processing steps mainly consist of filtering, demultiplexing, and adaptor trimming, which is preceded by a quality check of the sequenced reads by tools such as NGS QC toolkit [8] and FastQC [9]. These tools generate a well-structured report and provide complete information regarding the FASTQ files. Depending upon the fastQC report, filtering of the reads is done. Reads in the FASTQ files are filtered out based on their base call quality (Phred score) and the read length. The filtering step reduces the detection of false-positive variants as they have poor confidence base calls. Also, very short reads may hamper the mapping process as they may align to multiple regions in the genome. In NGS, multiple samples are sequenced simultaneously in the same instrument. Hence, demultiplexing is performed to separate the sequencing reads belonging to a particular sample using a unique barcode assigned to individual samples. Finally, adaptor trimming removes the library adaptor sequences from the ends of the demultiplexed reads. This step ensures that they do not interfere with the mapping and assembly processes. Trimmomatic [10] and Cutadapt [11] are the two most commonly used tools for this analysis.

The secondary analysis mainly encompasses the alignment of the reads against a reference human genome, or a de novo assembly, followed by variant calling. In sequence alignment, reads are aligned against a known reference genome, e.g. hg19 or hg38 for humans, thereby determining the genomic coordinates of the read. BWA [12], Bowtie [13], minimap2 [14], Magic-BLAST [15] are a few mapping tools used for read-alignment. De novo assembly is based on graphs theory, where reads are aligned to each other based on their sequence similarity, and no reference is used. SAM/BAM files are generated after the sequence alignment. SAMtools [16] are used to manipulate these files. Once the reads are aligned, three intermediate steps need to be performed before variant calling: duplicate removal, local read alignment, and base quality recalibration. During the library preparation, the Polymerase Chain Reaction (PCR) technique generates duplicate reads, which may cause false positives. Hence, they are removed from the analysis using Picard tools (http://broadinstitute.github.io/picard/). The presence

of InDels may cause read mismatch; hence, local read alignment is used to reduce this mismatching. Phred-scaled quality score, generated by the sequencers, may get affected by factors such as the sequencing platform and the sequence composition and thus may not reflect the actual base-calling error rate. Therefore, it is important to recalibrate the base quality score to enhance variant calling accuracy. Genome Analysis Toolkit (GATK) [17] is used for local read alignment and base quality score recalibration. Variants are then identified in the post-processed BAM file using variant callers, relying on Bayesian approaches, likelihood or machine learning algorithms that have significantly evolved over recent years. Most variant callers generate a variant calling format (VCF) file as their output. Variants identified may range from single nucleotide variants (SNVs) and INDELS (insertions and deletions) to complex chromosomal aberrations such as translocations, inversions, and copy number gains or losses (CNVs). Tools such as varscan2 [18], MuSE [19], Mutect2 [20] and SomaticSniper [21] are used to identify SNVs and INDELS from WES/WGS data. Delly [22], BreakDancer [23] and Pindel [24] identify CNVs and structural variants (SVs) from WGS data.

The final step of the NGS analysis is the data interpretation, i.e. determining the association between variants detected and the phenotype observed in a patient. The tertiary analysis involves variant annotation and variant filtering followed by data visualization. Variant annotation ascertains the biological or functional impact of the genetic variants in addition to providing the variant context. Variant Effect Predictor (VEP) [25], Annovar [26], snpEff [27] are the most widely used annotation tools. These tools use the vcf files obtained from the variant callers and provide information such as the chromosomal location of the variants and the biological impact of the variants, i.e. if the variant is missense, nonsense, synonymous, stop-gain, stop-loss, etc. Variants are filtered out based on their impact, thereby increasing the probability of detecting an actionable/driver variant. When we want to discover disease-causing rare variants in the data, variants commonly found in the population, i.e. SNPs (Single Nucleotide Polymorphism), can be removed from the analysis using databases such as dbSNP [28]. SNPs are the single nucleotide variants found in at least 1% of the population. Multiple scores such as SIFT[29], Polyphen[30], FATHMM-XF[31] and CADD[32] remove the benign variants from the analysis. Further, population databases like COSMIC[33], ClinVar [34] and OncoKB [35] are used to determine the clinical association of variants. After retrieving the final set of variants, they are correlated with the phenotypic characteristics of the patients. If there is a significant association between them, the findings are further validated biologically. For example, we can analyze the genomic data at multiple time points to examine the change in the variants and if any of those variants are linked to disease progression in the patients. These analyses also assist in developing effective drugs for specific diseases. They are particularly helpful in personalized cancer therapy, where a subject's response to a particular drug can be captured. Thus,

determining the correlation of the drug with the mutation profiles of the subject can aid in designing targeted medicine. A plethora of meaningful information is derived

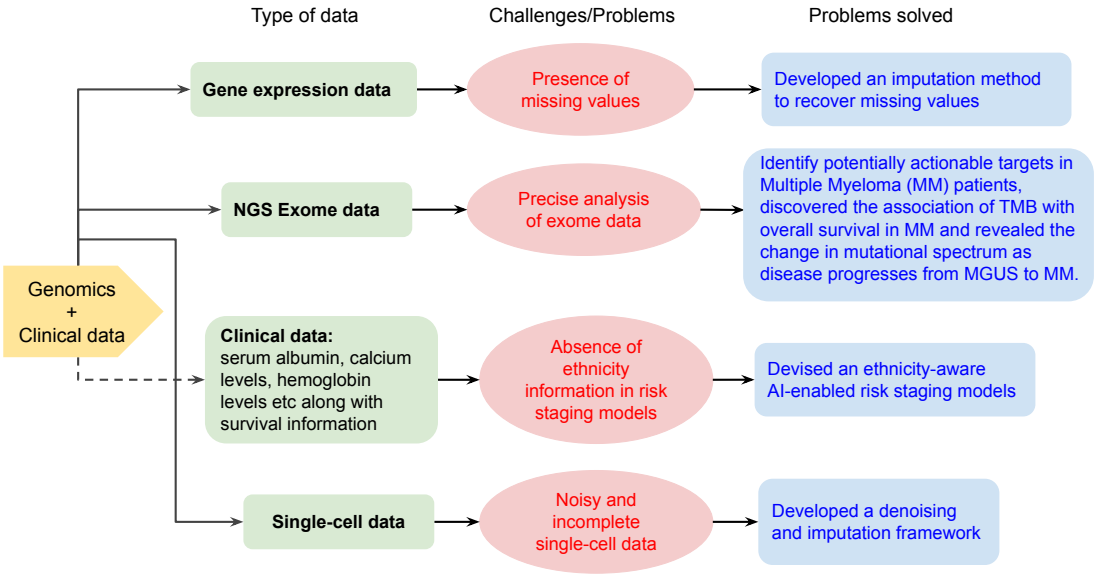


Figure 1.2: Different challenges that exist in cancer genomics. Development of computational strategies to tackle these problems.

from the sequencing data, such as identifying upregulated and downregulated genes, detecting genomics variants, copy number variants, or chromosomal alterations. Diverse computational methods are being developed to analyze genomic data. However, sometimes the data gets corrupted during the acquisition process and may contain noise in the form of missing values, false positives and false negatives. Hence, we need to devise methods to address such challenges. For example, gene expression data derived from molecular techniques may include missing values, and subsequent analysis of this incomplete data may lead to inaccurate findings. Not only this, single-cell data analysis has now gained preference over bulk data owing to its better resolution but it is often corrupted by the presence of false positives, false negatives, and missing bases. Thus, given the significance of using noise-free data for the precise interpretation of the findings, we have formulated and proposed different computational solutions to deal with problems in cancer genomics in this thesis. While working with multiple myeloma data in imputation problem, we became interested in pursuing research in MM. Another contributing factor was the availability of exome and clinical data of MM. MM is a blood cancer characterized by abnormal growth of plasma cells in the bone marrow. MM is preceded by a benign condition of Monoclonal Gammopathy of Undetermined Significance (MGUS). We analysed the exome data of MM patients to unravel the heterogeneity present in MM patients. We also analysed the exome data of MGUS patients to determine the factors responsible for the disease progression to MM. Overall survival in MM varies depending upon the tumor heterogeneity and the initial risk stage of the

patient as these are the deciding factors for the therapy to be given to the patient and the subsequent treatment response. Furthermore, disease biology is also impacted by the ethnicity of the patient. Therefore, we came up with the idea of devising a machine-learning (ML) based method ethnicity aware method for risk stage prediction in MM patients. The various challenges addressed in the thesis and the different types of genomic data handled are shown in the Figure 1.2.

1.2 Literature review

High dimensional gene expression data is crucial in studying the relationship between genes and diseases including cancer. Molecular techniques such as 'Microarrays' facilitate estimation of expression levels of thousands of genes simultaneously under different experimental conditions and gene expression data generated from such experiments is subsequently analyzed using statistical or machine learning methods to extract relevant information. Microarray data is useful in a wide range of applications starting from disease diagnosis to drug discovery. It aids in subject risk stratification, classification of clinical subtypes and prediction of response to therapy. These analyses further assist in developing effective drugs as the treatment strategies are targeted directly to the specific type of cancer. subject's response to a particular drug may be captured and correlation between therapeutic responses to drugs and the genetic profiles of the subjects can be evaluated leading to personalized medical treatment.

A persistent problem associated with microarray dataset is the presence of varying number of missing values in the data that may arise owing to poor slide quality (dusty or scratchy), poor image quality, or insufficient resolution [36]. Subsequent downstream analysis on incomplete gene expression matrices may be highly inaccurate. One of the ways of dealing with the problem of missing values is to capture microarray data again but it does not guarantee complete data matrix. Moreover, the entire process is expensive and time consuming. An alternate solution to this problem is to remove the genes containing missing values from the analysis. However, this can result in loss of information and may lead to inaccurate findings on driver genes and/or altered biological pathway. Therefore, it is worthwhile to apply advanced computational methods for the imputation of missing values in microarray data prior to any analysis.

Numerous methods have been developed in the recent times for imputation of gene expression data. These can be broadly categorized into four classes: hybrid methods, local methods, global methods, and knowledge assisted methods (Table 1.1). Some of the early methods developed to account for the missing values are ZEROimpute, ROWimpute and COLimpute [37]. In ZEROimpute, missing values are replaced with zeros.

In ROWimpute and COLimpute, missing values are replaced with the averaged values of the observed entries of the corresponding rows or columns. These methods do not take into consideration the correlation present among genes and therefore, do not perform optimally. Gene expression matrix is highly correlated. Therefore, it is important to consider correlation among genes. Several methods exist in literature based on correlation among genes. These are categorized into local and global approaches based on the type of correlation utilized by them. As shown in Table 1.1, local approaches impute missing values by considering the group of genes that show high correlation with the gene containing missing values. Such methods perform optimally when the data is heterogeneous. *k* nearest-neighbor imputation (KNNimpute) [38] is one of the earliest local approach method to impute missing value. It first estimates *k* nearest group of genes that are similar to the missing target gene, followed by averaging of these genes to impute the missing value of the target gene. SKNNimpute (Sequential KNNimpute) [39] and IKNNimpute (iterative KNNimpute) [40] are variations of KNNimpute. Gaussian mixture clustering imputation (GMCimpute) [41], least square imputation (LSimpute) [42] and variations to LLSimpute, sequential LLSimpute (SLLSimpute) [43], iterative LLSimpute (ILLSImpute) [44], robust least square estimation with principal components (RLSP) [45], Bayesian gene selection BGSregress [46], collateral missing value imputation (CMVE) [47] and auto-regressive least square imputation (ARLS) [48] are all examples of local approaches. On the other hand, SVDimpute (Singular Value Decomposition) [38], Bayesian Principal Component Analysis (BPCA) [49] are the examples of global approach and utilize the global correlation present in the entire gene expression matrix. Hybrid approaches include methods like LinCmb [50], HPM-MI [51]and tri-imputation [52]. GOimpute [53], HAimpute [36] and (iMISS) [54] are knowledge-assisted methods that combine the already existing domain knowledge to imputation techniques for imputing missing values in gene expression data, thereby, increasing their imputation accuracy. Gene Ontology based similarity measure has been recently used for missing value imputation in miRNA microarray data [55]. A brief review of all the existing methods is shown in Table 1.1.

Global approach based methods such as SVDimpute [38], Bayesian Principal Component Analysis (BPCA) [49] exploit the global covariance information resulting from the entire gene expression matrix. In SVDimpute, singular value decomposition is used to calculate principal components of gene expression matrix, referred to as eigengenes, which can then be linearly combined to approximate the expression of all genes in the data set. SVDimpute first performs linear regression of the target gene against the 'k' most significant eigengenes and then uses the coefficients of the regression to reconstruct the missing values from a linear combination of the 'k' eigengenes. BPCA [49] performs missing value imputation in three steps: a) Principal Component (PC) regression, 2) Bayesian estimation and 3) an Expectation-Maximization (EM)-like al-

gorithm. N-dimensional gene expression vectors are expressed as a linear combination of K principal axis vectors and an EM-like algorithm is then used to estimate the posterior distributions of the model parameter and the missing values simultaneously.

Table 1.1: Review of existing methods for missing value imputation in gene expression data

	Local Approach	**Global Approach**	**Hybrid Approach**	**Knowledge assisted Approach**
Method	Imputes missing values by first estimating the local correlation among the group of genes that are highly correlated with the gene containing missing values and then using the local correlation to calculate the missing value.	Imputes missing values by utilizing the global correlation among the genes in the complete gene expression matrix.	Exploits both the global and local correlation among genes to calculate missing values in gene expression data.	Imputes missing values by integrating already existing domain knowledge to imputation methods. Information about biological process in the microarray experiment etc. is an example of domain knowledge that can be integrated to the method.
Advantages	Perform optimally when the data is heterogeneous i.e genes exhibit dominant local similarity structure.	Perform optimally when the data has high global covariance in expression matrix.	Perform optimally regardless of the type of covariance present in the gene expression data.	Improves accuracy of missing value imputation and perform optimally in presence of noisy data.
Limitations	Perform poorly when data lacks local similarity structure.	Fail to perform well when the data is heterogeneous.	Perform sub optimally when data is noisy and has high missing rates.	Perform sub optimally when data has high missing rates.
Examples	(i) K nearest-neighbor imputation (KNNimpute) (Troyanskaya et al., 2001) and its variations- SKNNimpute (Sequential KNN) (Kim et al., 2004b), IKNNimpute (iterative KNNimpute) (Bŕas and Menezes, 2007) (ii) Gaussian mixture clustering imputation (GMCimpute) (Ouyang et al., 2004) (iii) Least square imputation (LSimpute) (Bø et al., 2004) and its variations- Local least square imputation (LLSimpute) (Kim et al., 2004a), Sequential LLSimpute (SLLSimpute) (Zhang et al., 2008), iterative LLSimpute (ILLSImpute) (Cai et al., 2006) and robust least square estimation with principal components (RLSP) (Yoon et al., 2007) (iv) Bayesian gene selection BGSregress (Zhou et al., 2003), Collateral missing value imputation (CMVE) (Sehgal et al., 2005), Auto-regressive least square imputation (ARLS) (Choong et al., 2009)	(i) Bayesian Principal Component Analysis (BPCA) (Oba et al., 2003). (ii) SVDimpute (Singular Value Decomposition) (Troyanskaya et al., 2001) first estimates principal components of gene expression matrix by calculating Singular value decomposition of the gene matrix and it then selects the most significant components. These selected components are further used to approximate missing values in the gene expression data.	(i) LinCmb (Jörnsten et al., 2005) uses both global and local correlation information in the data. It estimates missing values using five different imputation algorithms, row average, KNNimpute, GMCimpute, SVDimpute and BPCA. It then takes a convex combination of the results obtained from each of the methods to compute final result. (ii) HPM-MI (Hybrid Prediction Model with Missing value Imputation) (Purwar and Singh, 2015) is a hybrid approach that uses both k-means clustering and Multilayer perceptron. It uses eleven different missing value imputation techniques to compute missing values and then selects the best clusters using k-means to compute final result. (iii) Tri-imputation (He et al., 2016) employs three base imputation algorithms to impute the genes with missing values.	(i) GOimpute (Tuikkala et al., 2005) uses the prior information about the functional similarities in term of GO for missing value imputation. (ii) HAimpute (Imputation using Histone Acetylation information) (Xiang et al., 2008) combines histone acetylation information as domain knowledge with imputation methods such as KNNimpute and LLSimpute. Accuracy of missing value imputation improves considerably after utilizing domain knowledge.

Most of the methods perform missing value imputation in gene expression data at comparatively higher observability, say, when 70% or more data is available (that is equivalent to 30% or less data is missing). Recent developments have made it possible to predict expression data values when the observed data is as low as 10%. Gene expression data is a highly correlated data because of the high level of interdependence between the genes. This interdependence is due to functional relationship between the genes as the group of genes interact together in any biological process. Therefore, it is evident that gene expression matrix is very similar to a low rank matrix that can be embedded into a lower dimensional subspace. Imputation of missing values in data matrix can be projected as the matrix completion problem and hence, we devised an optimisation based framework for imputation which is explained in detail in chapter 2. The proposed method has been tested on datasets of hematological malignancies. While working with multiple myeloma (MM) dataset on imputation problem, we decided to focus on only one type of blood cancer i.e. MM. The problem of missing value imputation is a generic one and can be easily applied to datasets of different cancer types. There has been tremendous amount of research in multiple myeloma, but there are gaps remaining that need to be addressed. However, such type of research work requires access to the longitudinal exome and clinical data of cancer patients. We were lucky to have the availability of the exome data and clinical data of multiple myeloma from our collaborators and authorized access to online datasets which further motivated us to pursue this problem.

Multiple Myeloma (MM) is a malignancy of clonal plasma cells that tend to evolve and accumulate as disease progresses from precursor transition states of Monoclonal gammopathy of undetermined significance (MGUS)/Smoldering Multiple Myeloma (SMM) to active MM and ultimately Extramedullary disease/Plasma cell leukemia (PCL). Reservoir founder clones may exist prior to MGUS [56], that may become detectable and dominant with progression and gradually evolve into heterogeneous subclones. The process of subclonal propagation of plasma cells (PCs) during myelomagenesis is complex and is driven under the influence of selection pressures exerted by immune surveillance, microenvironment and therapeutic agents.

Molecular mechanisms that underlie early progression in newly diagnosed MM patients who fail to respond to existing treatments are not completely understood. MM shows heterogeneity in terms of clinical phenotypes, rates of disease progression, response to therapy and survival outcomes, all of which are influenced by the underlying genomic complexity of the patient [57]. It is established that two types of primary oncogenic events are involved in initiation of myelomagenesis [58, 59] as shown in Figure 1.3. These include IgH translocations (found in nearly 55% patients) and hyperdiploidy of odd numbered chromosomes (observed at a frequency of 40%). These two kinds of aberrations may coexist in nearly 10% of cases. A gamut of secondary events (muta-

tions in RAS, NF-κB pathway, overexpression of MYC, haploinsufficiency of p53, (1q) gain and (1p) loss) are known to occur that provide further growth advantage to evolving (sub)clones, promote drug resistance, genome instability and progression. Deletion 13q is commonly found among non-hyperdiploid MM as well as in MGUS which suggests its role as a primary event during early oncogenesis of MM [60, 61, 62].

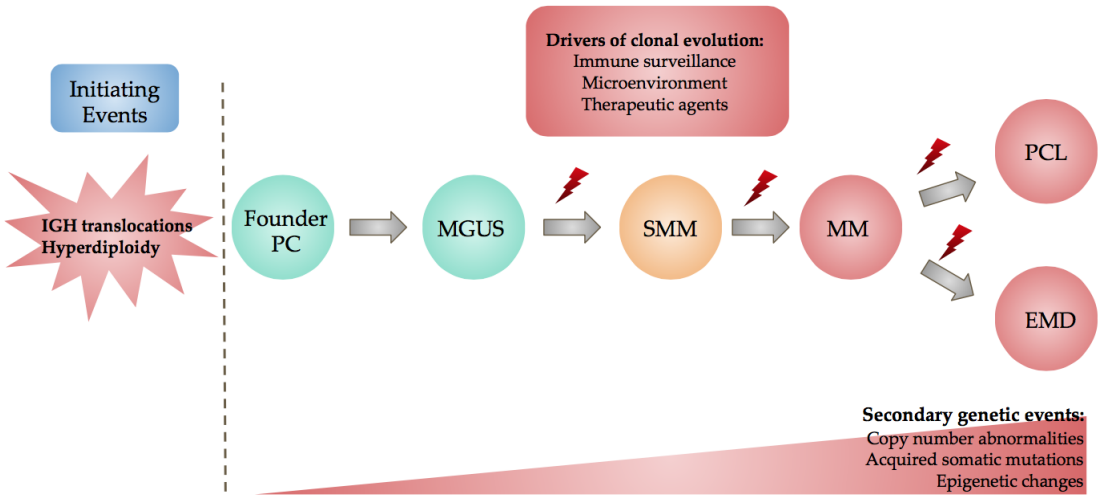


Figure 1.3: Clonal evolution in MM. MM is initiated by events like IGH translocations or Hyperdiploidy. MGUS and SMM are the precursor stages of MM which progresses to MM over time on acquisition of multiple genetic changes, ultimately leading to PCL and EMD.

Based on mutational complexity and subclonal architecture, different patterns of clonal evolution have been reported in MM. The branching type of clonal evolution analogous to Darwinian model is the most frequent one and is found in $\geq$ 50% MM patients whereas linear or stable evolution with no significant alteration in subclonal architecture have been observed in $\leq$ 30% cases [63, 64, 65]. Analysis of WES data obtained from MM patients on Immunomodulatory imide drugs (IMiDs) from UK Myeloma XI phase 3 trial and the CoMMpass study has revealed that 20% MM patients experienced neutral tumor evolution associated with poor prognosis while remaining 80% encountered branching evolution [11]. Patients with branching evolution may respond well to IMiDs as these can reconfigure bone marrow stromal cum immune microenvironment and prolong survival [11]. Instead, patients with neutral clonal evolution with random genetic drift may benefit from combinations of PIs with high dose melphalan [66, 67].

Recent NGS studies conducted on pairwise myeloma genomes/ exomes at two or more serial time points have reported presence of intraclonal heterogeneity during progression and relapse [56, 62, 63, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76]. A series of somatic mutations including substitutions, indels and copy number variations emerge during disease progression that contour the pattern of clonal evolution. Numerous driver mutations have been identified in myeloma genome [72] that may co-evolve mutually in

cooperation or exclusively either in same or different (sub)clones and modulate their net impact on clinical outcomes.

Although clonal heterogeneity in MM is well established, subclonal remodelling of gains/ losses and rewiring of functional pathways are not completely understood. There is currently a paucity of data available on longitudinal subclonal evolution profiles associated with progression in MM and a deeper understanding is required to assess mutations of clinical relevance that could potentially be targeted for treatment in future therapeutic approaches against MM and its precursor states [66, 77]. The progressing subclonal shifts are of paramount clinical significance as these could promote oncogenesis and lead to drug refractoriness. Estimation of their cellular prevalence could further predict likelihood of depth of response and a rationalized approach of combinatorial therapy. More and more longitudinal studies are needed to explore the progressing subclonal events and ultimately guide combinations of targeted therapy that can eradicate such subclonal populations and delay progression. Hence, we decided to conduct this study to capture subclonal mutational landscapes associated with progression of MM and identify potential actionable/ druggable targets that can be treated with their corresponding drugs. This work is explained in detail in chapter 3. Multiple myeloma is preceded by a non malignant condition, Monoclonal Gammopathy of Undetermined Significance (MGUS). In MGUS, there is normal as well as abnormal plasma cells in the body. The abnormal plasma cells lead to the presence of M protein in the blood. MGUS does not usually cause any clinical symptoms and often go undetected till it transforms to full blown MM. Since, the tumor initiating genomic alterations have already taken place at MGUS level, it is important to study MGUS and identify prognostic factors responsible for the malignant transformation of MGUS to MM.

Monoclonal gammopathy of undetermined significance (MGUS) is a benign precursor state of MM characterized by lack of end-organ damage [78] and less than 10% of plasma cells in the bone marrow. MGUS may progress to asymptomatic or symptomatic multiple myeloma with a rate of nearly 1% per year [79], where MM is characterized by severe clinical problems such as bone fractures, anaemia, renal failure, and hypercalcemia. Multiple studies involving exome and genome data of MM have been performed to understand the genomic abnormalities driving tumor progression in MM. It is well established that the primary events in MM are either hyperdiploidy, i.e., trisomy of chromosomes 3,5,7,9,11,15,19 and/or 21 or non-hyperdiploidy involving translocations affecting the genes encoding immunoglobulin (Ig) heavy chains (IGH)-mainly t(4;14), t(6;14), t(11;14), t(14;16) and t(14;20) [80]. Primary events are then followed by multiple secondary events promoting tumor progression. However, it has also been observed and validated that the genetic aberrations peculiar to MM are also present during the premalignant state of MGUS, where they do not show any clinical symptoms related to MM [73, 81]. It is, therefore, worthwhile to thoroughly investigate the mutational

landscape of the genomic alterations affecting MGUS as well as MM. Though multiple studies have been performed to study the MGUS to MM progression [72, 82, 83], the landscape of the mutational patterns of the MGUS and MM largely remains unexplored. The study of the changing mutational spectrum of the MGUS as it advances to MM will provide more insight into the disease biology. Further, it will help identify the clinically relevant vital biomarkers that can assist in controlling the progression of MGUS to MM.

Mutational signatures have emerged as critical biomarkers in cancer genomics, with profound pathogenic, prognostic, and therapeutic implications. Multiple mutational events occur in a tumor, while only a few of these mutations are actual drivers of cancer. However, exploring the entire landscape of coding and non-coding mutations helps reveal the mutational signatures characteristics of the specific cancer types. For example, CG>AT transversion is associated with lung cancer [84], and CG>TA is associated with skin cancer [85, 86]. Various mutational signatures have been discovered based on the 96 possible combinations of the single base substitutions and their trinucleotide contexts. These signatures are linked with the defects of DNA repair mechanisms, ageing, UV exposure, and others, thereby validating the role of the mutational processes in shaping the genomic continuum of each cancer type [87, 88, 89]. It will be interesting to explore the association of MGUS and MM to with any of these mutational signatures. Further, tumor mutational burden (TMB) has become a prominent biomarker of response to immunotherapy and is being explored for its association with overall survival, particularly in solid tumors. TMB is determined as the number of mutations identified per megabase. It has been observed that cancers with a high TMB load of greater than 10 mut/Mb have a better chance of responding to drugs called immune checkpoint inhibitors (ICIs). The primary function of ICIs is to activate the immune system better to recognize cancer cells [90] and act upon them. As a result, a high tumor mutational burden (TMB) has been increasingly associated with superior overall survival in ICI-treated patients. Multiple studies are now being conducted to discover the cancers with high TMB that respond best to ICIs and, thus, prolong the survival of cancer patients. In addition, the association of TMB with survival in non-ICI-treated patients has also been explored. It has been observed that high TMB was associated with poor prognosis and overall survival in the absence of immunotherapy, as opposed to ICI-treated patients in whom high TMB was associated with prolonged survival [91]. Since, MM may have high mutation burden in certain patients, it will worthwhile to infer the association of TMB with the survival outcomes in MM.

Synonymous mutations, earlier designated as silent mutations, were mostly ignored in cancer genomics due to their inability to alter the amino acid of the resultant protein [92]. However, they have the capability of changing the protein expression and function owing to their impact on RNA stability, RNA folding [93] or splicing [94], translation

[95], or co-translational protein folding. Multiple studies have corroborated that natural selection is present in synonymous mutations [95, 96, 97], contrary to earlier studies that denied the role of selective pressure in synonymous mutations [98]. Various genome-wide association studies conducted in recent times have also confirmed the association of synonymous SNPs to human disease risk and other complex traits. Therefore, the role of synonymous mutations in the disease biology of MGUS and MM should be examined as it could lead to significant prognostic and clinical implications. Given the existing gap in the literature of MGUS and MM, we were inspired to explore and compare the evolving mutational spectrum as disease progresses from MGUS to MM.

In MM, the overall survival period ranges from 6 months to more than 10 years. The variability in the outcome of patients is an implication of the clinical and biological heterogeneity underlying MM. Substantial advances in tumor biology have made it possible to dissect the tumor heterogeneity present in MM, optimize patient treatment, and examine patient outcome. Multiple prognostic systems [99, 100, 101, 102, 103] have been described in MM that stratify patients into different risk groups. These risk groups further assist in identifying high risk patients who may require intense therapy upfront and/or a higher monitoring frequency during the follow-up periods. The first staging system for MM was proposed in 1975 [99] followed by the development of International staging system (ISS) [100] in 2005 and a Revised ISS (R-ISS) [101] in 2015. ISS utilizes serum albumin and β2-microglobulin while R-ISS makes use of ISS, Lactate dehydrogenase (LDH) and high-risk cytogenetic aberrations (HRCA). Currently, triplet combination therapy is the new standard-of-care in MM which has shifted many high risk patients to standard risk category, thereby justifying the need for a new risk-stratification system with the possibility of inclusion of more prognostic factors. Although human physiological and genetic profile is known to vary across ethnic groups, the current MM risk-staging systems do not account for ethnicity-specific information that can have a huge impact on the risk score prediction. It is evident from the studies that African Americans experience 2-3 times higher incidence rates than Asians, Mexican-Americans or Europeans [104]. Recent studies have observed significant variation in the overall survival of different groups belonging to distinct races/ethnicities since the introduction of novel treatment agents in MM [105, 106, 107, 108].In a recent study, vitamin-D deficiency at diagnosis was found to be a predictor of poor overall survival in MM [109]. However, this was significant only for White Americans and not for African Americans even at lower cut-offs of deficiency [109]. Similarly, HRCA, which is used to determine the intensity of frontline therapy, does not track with survival outcomes in African Americans10, thereby, highlighting the need for a race-specific risk-stratification system. Though ethnicity is an important prognostic factor in predicting the risk for MM [110], yet the variations in the clinical characteristics among the different ethnic groups have not been evaluated adequately. Therefore, it is desirable to

have a staging system that includes the variations in the clinical characteristics of the patients pertaining to distinct ethnic groups. In addition, it should be based on clinical and laboratory parameters that are easily accessible in healthcare settings across the globe.

Technology is constantly improving and so are the methods to sequence the genomic data. Single-cell technology has gained momentum in recent years as it has become more accessible to researchers due to reduction in library preparation and sequencing cost. It also provides higher efficiency as compared to bulk sequencing data. Single-cell vs bulk sequencing could be explained in simple terms by the following analogy. If we look at stars with bare eyes, we cannot focus on individual stars. However, if we look at stars with telescope, we get a clear picture of individual stars. Similarly, single-cell technology improves resolution as the focus is individual cells unlike group of cells in bulk data. For the same reason, single-cell technology is revolutionizing the field of cancer genomics as it uncovers the cellular heterogeneity. However, there is a downside to this technology. Single-cell data is complex and often is corrupted with noise which needs to be tackled before performing any data analysis. Otherwise, it could result in inaccurate findings. The challenges associated with single-cell data inspired us to follow research in single-cell data as the last component of the thesis. Also, we were able to broadly cover the different technologies used in the study of genomic data in this thesis.

1.3 Thesis Contributions

The major contributions of the thesis are summarized below:

1. A novel two-stage method was proposed to recover missing values in gene expression data by utilizing the row and column sparsity of the gene expression matrix in the Discrete Cosine Transform (DCT) domain. The first stage is the compressive sensing (CS) based framework that utilizes DCT based sparsity of the gene expression matrix to recover missing values. The recovered matrix is then denoised in the second stage, where the low-rank property of the matrix is utilized. The significant contribution of this work is exploiting DCT based sparsity to impute missing value in gene expression data which has never been done before in the context of gene expression data. Further, the significance of the imputation was established by classification and biological pathway analysis. The proposed method was tested on CLL (Chronic lymphocytic leukemia), AML (Acute myeloid leukemia) and MM (Multiple Myeloma) datasets since the focus of the thesis was hematological malignancies. The problem of missing value imputation is generic and can be applied to datasets of other cancer types easily. So, we decided to concentrate on MM for our subsequent research problems. MM

is a hematological cancer that arises from malignant transformation and deregulated proliferation of clonal plasma cells (PCs) in bone marrow. The progression of disease is driven by multiple factors including immune surveillance, microenvironment and therapeutic agents. MM is highly heterogeneous where drug resistance can be seen in patients even those who initially showed a good response to the treatment. Not only this, conventional treatment therapy might not work for all the patients owing to the heterogeneity of the disease. Thus, longitudinal studies involving MM patients might help in shedding some light over the genomic events leading to disease progression and drug resistance.

2. Bulk sequencing data of multiple myeloma patients was analysed to unravel the potentially actionable targets in multiple myeloma (MM) and thoroughly examine the genomic landscape of clonal evolution in cancer patients. The major contribution of this work is the detailed analysis of the variants found to be mutated in the patients at two time points via exploration of the longitudinal exome data of 62 MM patients. This data was collected at two different time points, one at the time of diagnosis of disease and the other time when the disease has progressed in the patient. An ensemble-based approach was adopted to identify a more reliable set of variants in each patient. The study provided critical insights into the recurrent subclonal shifts in known drivers, oncogenic and tumor suppressor genes. This study has been explained in chapter 3. MM is a unique type of cancer which has a benign precursor stage known as Monoclonal Gammopathy of Undetermined Significance (MGUS). MGUS patients do not show any clinical symptoms, however, studies have shown the presence of genomic complexity in MGUS patients. Hence, it is worthwhile to examine the genomic data of MGUS in conjunction with MM to understand the differentiating factors leading to the transformation of MGUS to MM. Once we identify such biomarkers, it is possible to restrict the progression of MGUS to MM and thus prolong the survival of MGUS patients. However, it is not easy to get MGUS genomic data because the condition is non malignant and people usually go undiagnosed till it transforms to cancer. But, we were able to get access to 61 patients of MGUS and thus, decided to study MGUS with MM.

3. MGUS is a non-malignant condition identified by the presence of abnormal protein (M protein) in the blood. This condition does not cause any problem in itself, however, it poses an increased risk of developing MM with time and therefore, needs to be continuously monitored. MGUS being a precursor of MM, shows complex genomic landscape similar to MM and is an area of concern for the researchers. There have been multiple studies related to MGUS and MM, but they lack an in depth evaluation of the entire spectrum of mutations occurring in both MGUS and newly diagnosed MM (NDMM) patients. Hence, to fill the existing gap, exome data of MGUS and MM patients was evaluated in chapter 4 to investigate the change in the mutational spectrum as the disease transforms from the benign condition of MGUS to malignant MM. An exhaustive investigation of all the mutations was done by categorising them into three groups- synonymous, non-synonymous and others. A statistically significant change in the mutational

spectrum from MGUS to MM was found. There was a statistically significant increase in the frequency of all the variants as well as the TMB values from MGUS to MM. It was observed that 3' and 5' UTR mutations were more frequent in MM and might be responsible for driving MGUS to MM via regulatory binding site. The frequency of high TMB was low and was found to be associated with poor overall survival in newly diagnosed multiple myeloma patients. Survival data of the MM patients was also utilised to infer the association of survival outcome with single base substitutions and ABOPEC activity. The abundance of survival data for MM patients motivated us to explore this data further for risk staging in newly diagnosed MM.

4. Risk staging is a critical step in deciding the course of treatment for the patient and may impact the overall prognosis of the disease. Along other prognostic factors that are critical for designing risk staging system for cancers, ethnicity based heterogeneity forms an integral part of it. Ethnicity is known to affect disease biology and hence, cannot be overlooked. Therefore, in chapter 5, an ethnicity-aware AI-enabled risk staging system for newly diagnosed multiple myeloma patients has been proposed. The model utilizes the parameters- age, albumin, β2M, albumin, calcium, eGFR, hemoglobin and information on cytogenetic abnormalities and ethnicity to predict the risk stage of any patient. The main contribution of the study is examining the impact of ethnicity on risk stage prediction and exploiting the ethnicity information for risk stage prediction in MM. The proposed method is robust and reliable and is better able to separate patients into different risk groups.

5. The thesis mostly deals with either bulk data or clinical data. Bulk data is merely a representative of group of cells and not individual cells which is why cellular complexity is often masked by bulk data. On the contrary, single-cell technology provides better resolution at cellular level and hence, has gained momentum in the last few years. Single-cell data provides cell-specific information and therefore, provides a detailed picture of the complexity and the heterogeneity present in tissues. Given the significance of single-cell data in cancer genomics, we wanted to explore this data to broadly cover the technologies used in the study of genomic data. Therefore, in chapter 6, an optimization-based method for denoising and imputing noisy and incomplete single-cell data has been devised to infer the pattern of clonal evolution from the imputed and denoised matrix. Single-cell data for multiple myeloma was not available so, we tested our data on other cancer datasets. The significant contribution of the work is the development of a robust and computationally fast method for single-cell data that can efficiently work on small-sized datasets and large-sized datasets.

1.4 Thesis Organization

Rest of the thesis is organized into different chapters. Chapter 2 is on missing value imputation in gene expression data. One of the persistent problems associated with gene expression data is the presence of missing values. Thus, we proposed an optimization based method, DSNN (Doubly Sparse DCT domain with Nuclear Norm minimization) for gene expression data imputation. In the first stage, missing values were recovered by formulating it as the CS-based reconstruction with double sparsity in the Discrete Cosine Transform (DCT). DSNN uses both column and row sparsity. The second stage was framed as the denoising problem and exploits the low-rank nature of the data matrix. The proposed method was compared with state-of-the-art methods.

In chapter 3, we studied the genomic landscape of clonal evolution in Multiple Myeloma using the whole exome sequencing data of 62 MM patients collected at two time points at AIIMS, New Delhi, first at the time of diagnosis followed by second instant on progression of MM to investigate the pattern of clonal evolution of MM in these subjects' data. A comparative evaluation of the variants at two time points along with an depth analysis of evolving founder clones revealed multiple driver mutations including those known to be actionable. The workflow and the main findings of the work are presented in the chapter while detailed analysis of individual patients is provided at the end of thesis. In addition, critical insights into the recurrent pattern of subclonal shifts in certain important genes is also presented.

Whole exome data of MGUS and MM patients was evaluated in chapter 4 to investigate the change in the mutational spectrum from MGUS to MM. An exhaustive investigation of all the mutations was done by categorising them into three groups- synonymous, non-synonymous and others. The critical findings of the study are presented in detail in the chapter along with the methodology followed. There was a statistically significant increase in the frequency of all the variants as well as the TMB values from MGUS to MM. It was observed that 3' and 5' UTR mutations were more frequent in MM and might be responsible for driving MGUS to MM via regulatory binding sites. Association of survival outcome with multiple prognostic factors is also presented in the study.

In chapter 5, an ethnicity-aware AI based method, Consensus based risk staging (CRSS), for risk stratification in MM was proposed. This method is based on easy to acquire clinical parameters like age, albumin, β2M, hemoglobin, eGFR, calcium and information of cytogenetic abnormalities and ethnicity. Method was validated on two different datasets, in-house dataset obtained from AIIMS (MMIn) and Multiple Myeloma dataset obtained from Multiple Myeloma Research foundation (MMRF). Its performance was remarkably better as compared to existing risk staging gold standard for Myeloma, i.e. Revised ISS (RISS) in terms of Kaplan-Meier curves, *p*-values obtained via Log-rank

test, hazard ratios and concordance index.

In chapter 6, ARCANE-ROG, Algorithm for Reconstruction of CANcer Evolution from single cell data using RObust Graph learning was proposed. ARACANE-ROG is an optimization based framework which denoises and imputes single cell data and infers the pattern of clonal evolution from the denoised single cell data. Method was extensively validated on multiple simulated datasets and real datasets. A comparative analysis of the proposed method with the state-of-the-art methods in terms of reconstruction error, False positive to False Negative (FPFN) ratio, Tree distance and V-measure revealed the robustness and efficacy of the proposed method. Our proposed method efficiently worked on small-sized datasets and large-sized datasets.

Chapter 7 summarizes the thesis work and provides suggestions for future work.

1.5 Publications from the Thesis

1.5.1 Journals

1. **Akanksha Farswan**, Anubha Gupta, Ritu Gupta and Gurvinder Kaur, "Imputation of gene expression data in blood cancer and its significance in inferring biological pathways", *Frontiers in oncology*, vol. 9, article no. 1442, pp. 1-14, January 08, 2020.

2. **Akanksha Farswan**, Anubha Gupta, Sriram K., Atul Sharma, Lalit Kumar, Ritu Gupta, "Does ethnicity matter in multiple myeloma risk prediction in the era of genomics and novel agents? Evidence from real world data", *Frontiers in Oncology*, pp. 1-14, November 09, 2021.

3. **Akanksha Farswan***, Lingaraja Jena*, Gurvinder Kaur*, Anubha Gupta, Ritu Gupta, Lata Rani, Atul Sharma, Lalit Kumar,"Branching clonal evolution patterns predominate mutational landscape in Multiple Myeloma", *American Journal of Cancer Research*, ISSN: 2156-6976, pp. 5659-5679 November 15, 2021. (*Authors contributed equally)

4. **Akanksha Farswan**, Anubha Gupta, Lingaraja Jena, Vivek Ruhela, Gurvinder Kaur, and Ritu Gupta,"Characterizing the mutational landscape of MM and its precursor MGUS", accepted in *American Journal of Cancer Research*, Elsevier, March 03, 2022.

5. **Akanksha Farswan**, Ritu Gupta, and Anubha Gupta, "ARCANE-ROG: Algorithm for Reconstruction of Cancer Evolution from single-cell data using Robust Graph Learning", *Journal of Biomedical Informatics*, Elsevier, vol. 129, May, 2022.

1.5.2 Posters

1. Gurvinder Kaur, Anubha Gupta, **Akanksha Farswan** Ritu Gupta, and Sriram K, "Inferring Biological Pathways in Multiple Myeloma after Missing Value Imputation", *Clinical Lymphoma, Myeloma and Leukemia*, 19(10), p.e67. (Presented at 17th International Myeloma Workshop 2019 (*IMW* 2019), Boston, USA).

2. Ritu Gupta, Gurvinder Kaur, **Akanksha Farswan**, Lingaraja Jena, Anubha Gupta, Lata Rani, Lalit Kumar, and Atul Sharma, "Clonal evoution in multiple myeloma evaluated by Whole Exome Sequencing", *Clinical Lymphoma Myeloma and Leukemia*, 21, p.S64. (Presented at 18th International Myeloma Workshop 2021 (*IMW* 2021), Vienna, Austria).

3. Gurvinder Kaur, Ritu Gupta, Lingaraja Jena, **Akanksha Farswan**, Anubha Gupta, Lalit Kumar, Lata Rani, and Atul Sharma, "Whole Exome Sequencing provides novel insights in synonymous and non-synonymous mutational landscapes of Multiple Myeloma", *Clinical Lymphoma Myeloma and Leukemia*, 21, pp.S65-S66. (Presented at 18th International Myeloma Workshop 2021 (*IMW* 2021), Vienna, Austria).

4. **Akanksha Farswan**, Anubha Gupta, Vivek Ruhela, Lingaraja Jena, Gurvinder Kaur, Sriram K, Ritu Gupta, Prognostic value of TMB and its association with overall survival in newly diagnosed multiple myeloma patients, Presented (online) at CMMC Symposium, Sept 26-28, 2021, Cologne, Germany.

Chapter 2

Missing value imputation in gene expression data

2.1 Introduction

Matrix completion is a popular and challenging area of research in various domains. It is evident from the literature review done in chapter 1, section 1.2 that there is a need for a robust method for imputing missing values in genes expression data. Such method should be able to recover missing values efficiently at low as well as higher observability of the data. Therefore, a novel 2-stage method, DSNN (Doubly Sparse DCT domain with Nuclear Norm minimization), has been proposed in the study for predicting missing values in gene expression data using Compressive Sensing (CS) based formulation. In the first stage, missing values were recovered in gene expression data by formulating it as the CS-based reconstruction with double sparsity in the Discrete Cosine Transform (DCT). Matrix obtained in first stage is considered a noisy version of the original/true matrix. Therefore, in Stage-2, denoising of the matrix recovered from Stage-1 is done by utilising nuclear norm minimization. It exploits the low rank property of the data matrix. Missing value imputation was performed on four blood cancer dataset at different observability of data (10% to 90%) using NMSE as evaluation metric. Significance of imputation was validated by two experiments. In the first experiment, classification of normal versus cancer subjects was carried out. In the second experiment, biological significance of imputation was ascertained by first identifying top 500 genes using SPARROW algorithm [111], followed by KEGG analysis on these top 500 genes. SPARROW (SPARse selected expRessiOn regulators identified With penalized regression) algorithm finds candidate tumor drivers from the 'selected expression regulators' (SERs). It defines SERs as the genes that drive dysregulated transcription leading to carcinogenesis. This algorithm regresses the gene expression values on the candidate SERs and provides a rank to each SERs based on the genes expression values of the samples. The method has been described briefly in Section 3. Once the ranking was done by SPARROW, top 500 ranked genes from the list were further studied by KEGG [112, 113, 114] using a web based application, Enrichr, developed and maintained by [115] and [116]. Many matrix completion methods exist in the literature and out of these methods, LMaFit [117], LogDet (Logarithm determinant) [118], and Robust PCA [119] are three different state-of-the-art matrix completion methods. The proposed method has been compared with these methods. LMaFit is based on matrix factorization, while LogDet implements nuclear norm minimization. RPCA performs

feature reduction and is quite robust to outliers. However, these methods have some limitations. LogDet becomes computationally expensive as the size of the matrix increases. LMaFit and RPCA-GD provide good performance, but their parameters need to be tuned properly for better recovery of missing values. Recently Kapur et al. [120] has used low rank constrained matrix completion method for imputing missing values in genomics.

2.2 Materials and Methods

2.2.1 Dataset Description

Four publicly available microarray gene expression dataset of different cancer types and different population have been used. Dataset-1 is Chronic lymphocytic leukemia (CLL) cancer dataset (GSE50006) from USA. CLL dataset contains expression values of 220 subjects across 54675 probe-ids and consists of two classes depending on whether the subject has CLL or not. There are 188 tumor samples and rest 32 are normal samples. Dataset-2 is Acute myeloid leukemia (AML) cancer dataset (GSE9476)[121] from USA. It contains gene expression values of 64 subjects across 22283 probe-ids. Two classes are present in the data. Label '1' corresponds to person suffering from AML and label '2' corresponds to healthy subject. There are 26 tumor subjects and 38 healthy subjects. Dataset-3 is Multiple Myeloma (MM) cancer dataset (GSE47552) [122] from Spain. It contains gene expression data of 99 subjects across 33297 probe-ids. It has data from 20 subjects with MGUS, 33 with high-risk SMM, 41 with MM and rest 5 were healthy subjects. Dataset-4 is Multiple Myeloma (MM) cancer dataset (GSE125361) belonging to Indian population. It contains gene expression data of 48 MM subjects across 58341 probe-ids.

Data was pre-processed to convert probe-ids to gene symbols because gene versus sample information is required for SPARROW analysis. It was observed that several probe-ids showed same gene names. To overcome this problem, gene expression levels of the probe-ids corresponding to the same gene name were averaged and gene versus sample matrix was created. After pre-processing, CLL dataset had 220 samples with expression values of 23348 genes. AML dataset had 64 samples with expression values of 13650 genes. MM-Spanish dataset had 99 samples with expression values of 23307 genes. MM-Indian dataset had 48 samples with gene expression values of 33973 genes. Since the range of gene expression values was very high (of the order of 10^6) for the CLL dataset, data was log transformed to reduce its dynamic range and to ensure that the smaller values were not shadowed by the higher values during the missing data

recovery method.

$$\mathbf{X}_{\text{log-transformed}}(i,j) = log_{10}(\mathbf{X}_{\text{original}}(i,j) + 1) \tag{2.1}$$

Matrix imputation was carried out on the sample versus gene matrices. After matrix imputation, only tumor samples of both the dataset were used for SPARROW analysis.

Workflow pipeline of the proposed analysis is shown in Figure 2.1. First of all, pre-processing of raw data was done as described in the previous section. Next, missing value imputation was carried out on four blood cancer dataset at different observability of data using Normalized Mean Square error (NMSE) as evaluation metric. Significance of imputation was validated by two experiments. In the first experiment, classification of normal versus cancer subjects was carried out. In the second experiment, biological significance of imputation was ascertained using SPARROW algorithm [111] followed by KEGG analysis on the top 500 genes identified by SPARROW.

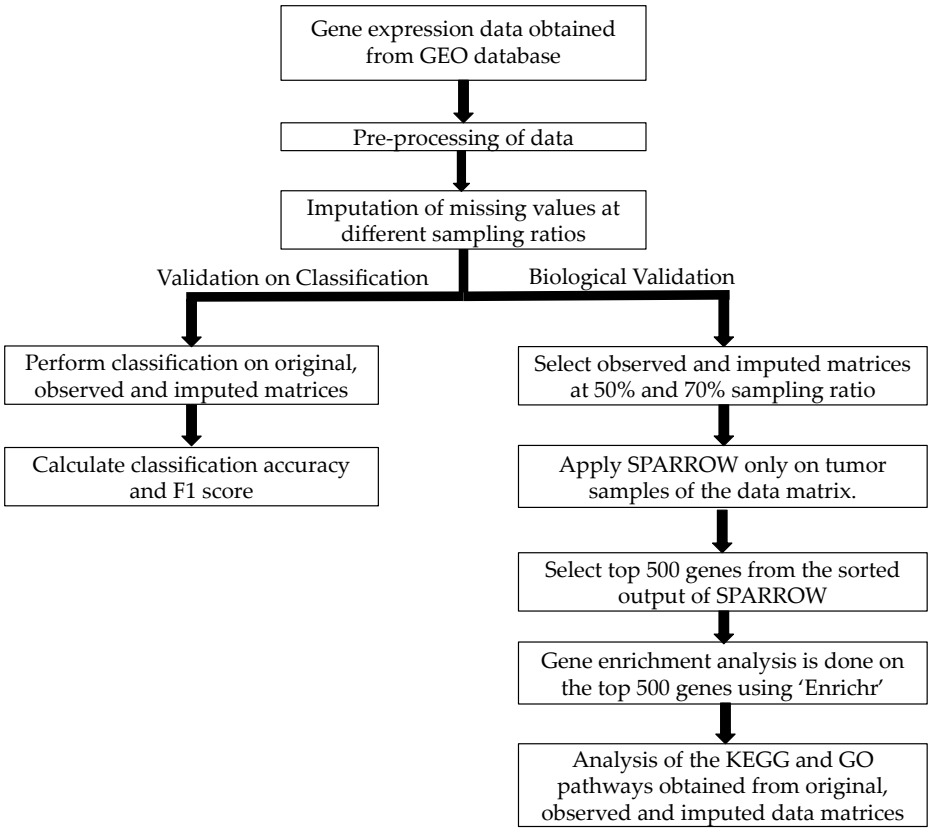


Figure 2.1: Workflow of the proposed analysis

2.2.2 DSNN Methodology

The proposed 'Doubly Sparse DCT domain with Nuclear Norm minimization' (DSNN) method consists of two stages. Stage-1 imputes missing values using a CS-based framework and DCT-based sparsity, while Stage-2 removes noise from the matrix obtained from Stage-1 by using a simple denoising framework.

Stage-1: Compressive Sensing based Matrix Completion: In this stage, missing value problem was projected as compressive sensing based reconstruction problem. To understand it better, consider an incomplete matrix $\mathbf{Y}$ of size $r \times s$, where r represents the number of subjects and s denotes the number of genes. Since the expression value of any gene will not vary much across subjects, data within a column would be sparse in some transform domain. Similarly, for a sample, gene expression levels of the gene will also be sparse in some transform domain. Columns and rows of the gene expression matrix were studied in the DCT domain and were observed to be highly sparse as shown in Figure 2.2. Based on this observation, Discrete Cosine Transform was chosen as the sparsifying transform in DSNN method because DCT acts as a KL-type basis for slow-varying signals [123] and data is sparse in the DCT domain.

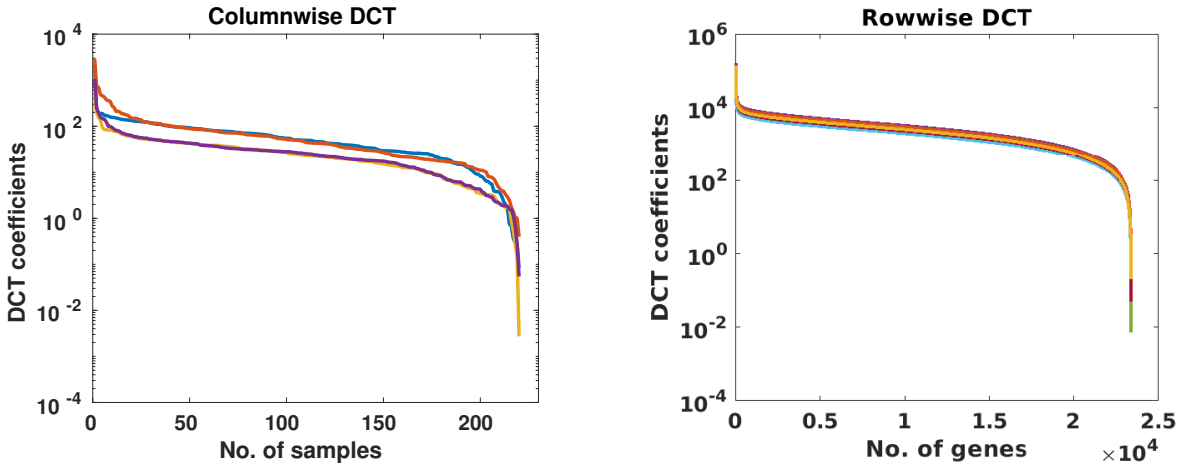


Figure 2.2: Each curve represents DCT coefficients of a few randomly chosen columns and rows of gene expression matrices of CLL dataset.

Thus, the missing data recovery problem was formulated in a compressive sensing framework, where the sensing matrix Φ was of size $r \times s$ and had '0' entries for missing values in data matrix $\mathbf{Y}$, while rest of the entries were '1'. Corresponding to each observed entry (that is not missing) of the i^{th} column, there is a row in Φ_i with an entry '1' for the corresponding position and zeros in the rest of the positions. For example, assume $\mathbf{x}_{\text{missing}}$=$\begin{bmatrix} x_1 & . & x_3 & . & . & x_6 \end{bmatrix}^T$ is the observed vector where only x_1, x_3 and x_6 are available and, x_2, x_4 and x_5 are missing (denoted as '.' in the vector). Then, the vector $\mathbf{x}_{\text{missing}}$ can be re-written as $\mathbf{y}$:

$$\mathbf{y} = \Phi\mathbf{x} \tag{2.2}$$

$$\mathbf{y} = \begin{bmatrix} x_1 \\ x_3 \\ x_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix}, \tag{2.3}$$

where the sensing matrix is written as $\Phi = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$ and $\mathbf{x}$ is the desired vector to be recovered. This is the standard formulation in compressive sensing literature, where it is assumed that only few values of data are sensed. In the above example, these values are x_2, x_4 and x_5. Thus, we have recast the problem of missing values in vector $\mathbf{x}_{\text{missing}}$ as the compressively sensed vector $\mathbf{y}$. Now, the task is to recover full data $\mathbf{x}$ from compressively sensed data $\mathbf{y}$ that will lead to missing value recovery.

Gene expression data was interpreted as a matrix with few observed samples, where the goal was to reconstruct the original/true matrix from the observed entries using DCT-based sparsity of gene expression data.

The following optimization problem was solved to recover the missing values in $\mathbf{Y}$

$$\underset{\tilde{\mathbf{X}}}{min}(||\mathbf{Y} - \Phi\tilde{\mathbf{X}}||_2^2 + \lambda_1||\mathbf{D}_c\tilde{\mathbf{X}}\mathbf{D}_r^T||_1), \tag{2.4}$$

where $\mathbf{D}_c$ is columnwise DCT matrix applied on columns of the $\tilde{\mathbf{X}}$ and $\mathbf{D}_r$ is the row-wise DCT matrix applied on rows of the $\tilde{\mathbf{X}}$. $\tilde{\mathbf{X}}$ is the matrix to be recovered. The above formulation is also known as analysis-prior and presence of DCT matrices in the formulation makes it non-separable. Using the orthogonal property of DCT transform, analysis prior was transformed to synthesis-prior formulation as

$$\underset{\mathbf{Z}}{min}(||\mathbf{Y} - \Phi\mathbf{D}_c^T\mathbf{Z}\mathbf{D}_r||_2^2 + \lambda_1||\mathbf{Z}||_1), \tag{2.5}$$

where $\mathbf{D}_c\tilde{\mathbf{X}}\mathbf{D}_r^T = \mathbf{Z}$. The above optimization problem was solved using the function handle and 'SPGL1' solver [124], [125], where the regularization parameter λ_1 was chosen automatically by the 'SPGL1' solver.

Stage-2: Denoising framework: It was assumed that the recovered $\tilde{\mathbf{X}}$ from Stage-1 is the noisy version of the original/true matrix $\mathbf{X}$ and hence, the recovered matrix was

denoised in Stage-2. Before denoising, $\tilde{\mathbf{X}}$ is re-organized into $\tilde{\mathbf{X}}_{\text{rec}}$ as

$$\tilde{\mathbf{X}}_{\text{rec}}(j,i) = \begin{cases} 0, & \text{if } (|\tilde{\mathbf{x}}(j,i) - \text{mean}(\mathbf{y}_i)| \geq \lambda_2 \text{std}(\mathbf{y}_i) \\ \tilde{\mathbf{x}}(j,i), & \text{otherwise} \end{cases} \tag{2.6}$$

where j ranges from 1 to m (number of rows/ subjects), $|.|$ denotes the absolute value and, mean($\mathbf{y}_i$) and std($\mathbf{y}_i$) denote the mean and the standard deviation of the i^{th} column of the initial observed (but incomplete) matrix $\mathbf{Y}$. Parameter λ_2 was determined empirically and was set to value 0.2 for experiments on CLL dataset, MM-Spanish dataset, and MM-Indian dataset. It was set to 0.1 for experiments on AML dataset. Denoising was formulated in the Split-Bregman type optimization as

$$\underset{\mathbf{W}}{min}\big(||\mathbf{W}||_* + \lambda_3||\mathbf{W} - \widehat{\mathbf{X}} - \mathbf{B}||_F^2\big) \; s.t. \; \widehat{\mathbf{X}} = \mathbf{W}, \tag{2.7}$$

where $\mathbf{B}$ is randomly initialized matrix and $\hat{\mathbf{X}}$ was initialized as:

$$\widehat{\mathbf{X}} = \tilde{\mathbf{X}}_{\text{rec}} + \tilde{\mathbf{X}}_{\text{inv-rec}} \circ rand(m,n), \tag{2.8}$$

where '$\circ$' represents the Hadamard product of two matrices with the elements of $\tilde{\mathbf{X}}_{\text{inv-rec}}$ defined as

$$\tilde{\mathbf{X}}_{\text{inv-rec}}(j,i) = \begin{cases} 1, & \text{if } \tilde{\mathbf{X}}_{\text{rec}}(j,i) = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2.9}$$

This step involves applying nuclear norm on the matrix $\mathbf{W}$ which is essentially the matrix that we want to recover. Nuclear norm is defined as the sum of the eigen values of a matrix as below:

$$||\mathbf{W}||_* = \sum_i \sigma_i(W) \tag{2.10}$$

Significance of using nuclear norm in the denoising framework is to recover low rank matrix here because gene expression data is low rank in nature owing to the interdependence between the different genes. Equation 2.7 was solved in Split Bregman type iterations as

$$\mathbf{W}^{k+1} = SVT_{\lambda_3}(\widehat{\mathbf{X}}^k + \mathbf{B}^k), \tag{2.11}$$

$$\mathbf{B}^{k+1} = \widehat{\mathbf{X}}^k + \mathbf{B}^k - \mathbf{W}^k, \tag{2.12}$$

$$\widehat{\mathbf{X}}^{k+1} = \tilde{\mathbf{X}}_{\text{rec}} + \tilde{\mathbf{X}}_{\text{obs}} \circ \mathbf{W}^{k+1}, \tag{2.13}$$

where 'SVT' denotes the soft singular value thresholding method [126] and $\tilde{\mathbf{X}}_{obs}$ is the observed incomplete matrix. Optimal value of parameter λ_3 was determined using grid search and was set to 100 in all experiments. All the randomly initialized matrices consist of uniformly distributed random numbers in the scale of 0 to 1. The complete algorithm for the proposed DSNN method is presented below.

Algorithm 1 Proposed DSNN Method

Stage 1 - Compressive sensing based matrix recovery

$\mathbf{Y}$ (Given incomplete matrix), ϕ, Discrete Cosine Transform matrices $\mathbf{D}_r$, $\mathbf{D}_c$

Obtain $\mathbf{Z}$ by solving $\underset{\mathbf{Z}}{min}(||\mathbf{Y} - \Phi\mathbf{D}_c^T\mathbf{Z}\mathbf{D}_r||_2^2 + \lambda_1||\mathbf{Z}||_1)$ using 'spgl' solver

$\tilde{\mathbf{X}} = \mathbf{D}_c^T\mathbf{Z}\mathbf{D}_r$

$\tilde{\mathbf{X}}$

Stage 2: Nuclear-norm based denoising

$\tilde{\mathbf{X}}$ (Recovered Matrix from Stage-1 considered as the noisy matrix)

$$\tilde{\mathbf{X}}_{\text{rec}}(j,i) = \begin{cases} 0, & \text{if } (|\tilde{\mathbf{x}}(j,i) - \text{mean}(\mathbf{y}_i)| \geq \lambda_2 \ \text{std}(\mathbf{y}_i) \\ \tilde{\mathbf{x}}(j,i), & \text{otherwise,} \end{cases}$$

$\widehat{\mathbf{X}} = \tilde{\mathbf{X}}_{\text{rec}} + \tilde{\mathbf{X}}_{\text{inv-rec}} \circ rand(m,n)$

$while$ converge:

$\mathbf{W}^{k+1} = SVT_{\lambda_3}(\widehat{\mathbf{X}}^k + \mathbf{B}^k)$

$\mathbf{B}^{k+1} = \widehat{\mathbf{X}}^k + \mathbf{B}^k - \mathbf{W}^k$

$\widehat{\mathbf{X}}^{k+1} = \tilde{\mathbf{X}}_{\text{rec}} + \tilde{\mathbf{X}}_{\text{obs}} \circ \mathbf{W}^{k+1}$

$end\ while$

$\widehat{\mathbf{X}}$ (Recovered Matrix)

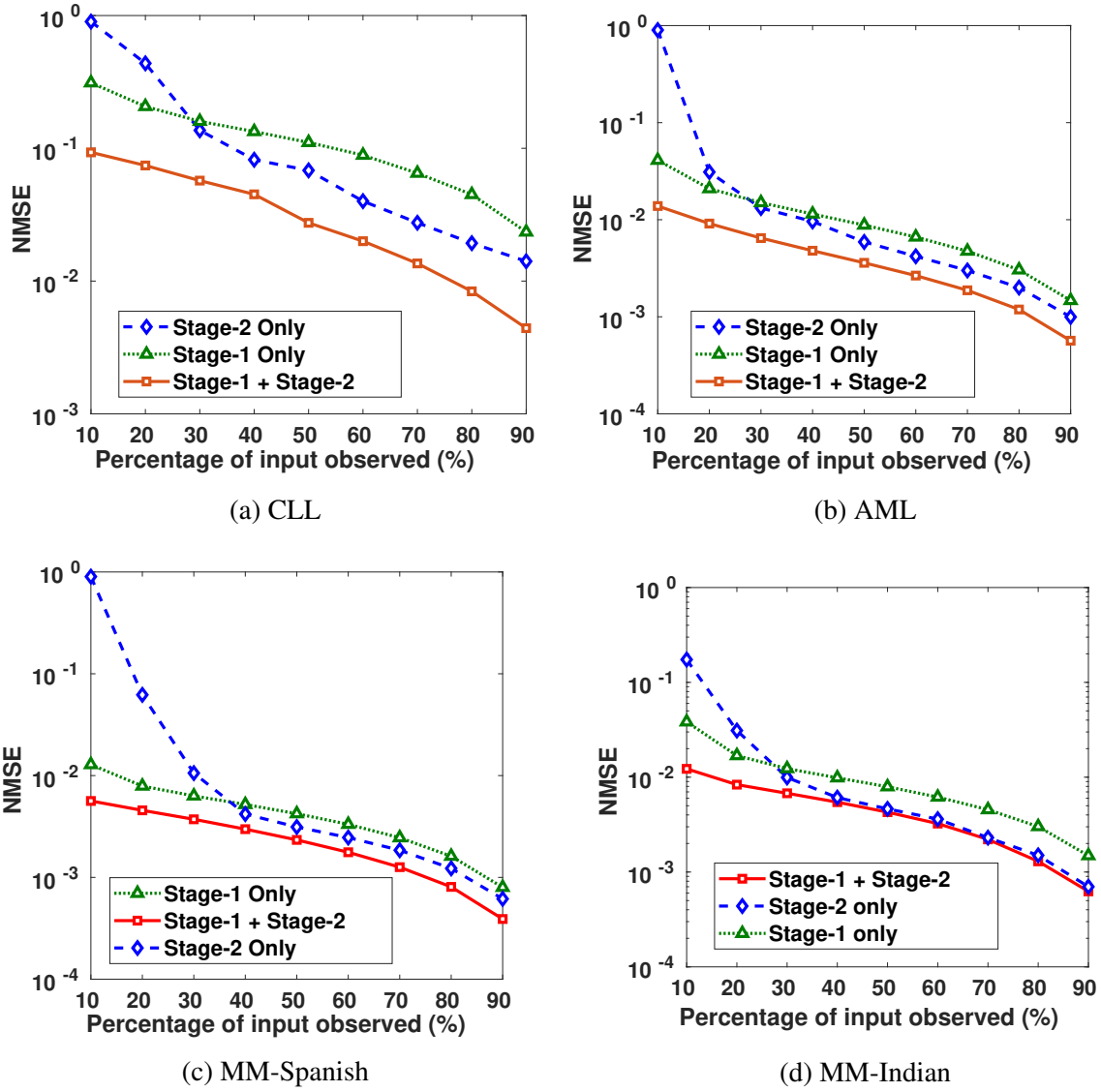
2.3 Results

2.3.1 Evaluation

For assessing the performance of the proposed DSNN method, some data were dropped randomly to create incomplete matrices with available data ranging from 10% to 90%. Next, incomplete matrices were imputed using the DSNN method. Results were simultaneously generated using three state-of-the-art matrix completion methods for comparative analysis. Normalized mean squared error (NMSE) was used as the evaluation metric and was calculated between the original/true and the recovered matrix. NMSE is defined as:

$$\text{NMSE} = \frac{||\mathbf{X}(original) - \hat{\mathbf{X}}(recovered)||_F^2}{||\mathbf{X}(original)||_F^2}. \tag{2.14}$$

Semi-log plots of NMSE at different stages are shown in Figure 2.3. Stage-1 results were obtained when missing values in data matrix were imputed using compressive sensing based matrix completion, where double sparsity in DCT domain was exploited. Stage-2 results were obtained when only nuclear norm minimization was used for matrix imputation. DSNN method combined both these stages. Results clearly indicated that the performance of imputation has improved with the two successive stages of DSNN. DSNN method also worked better than the existing methods even at high missing rates of 10% as shown in Figure 2.4. NMSE reported in the figures is averaged over

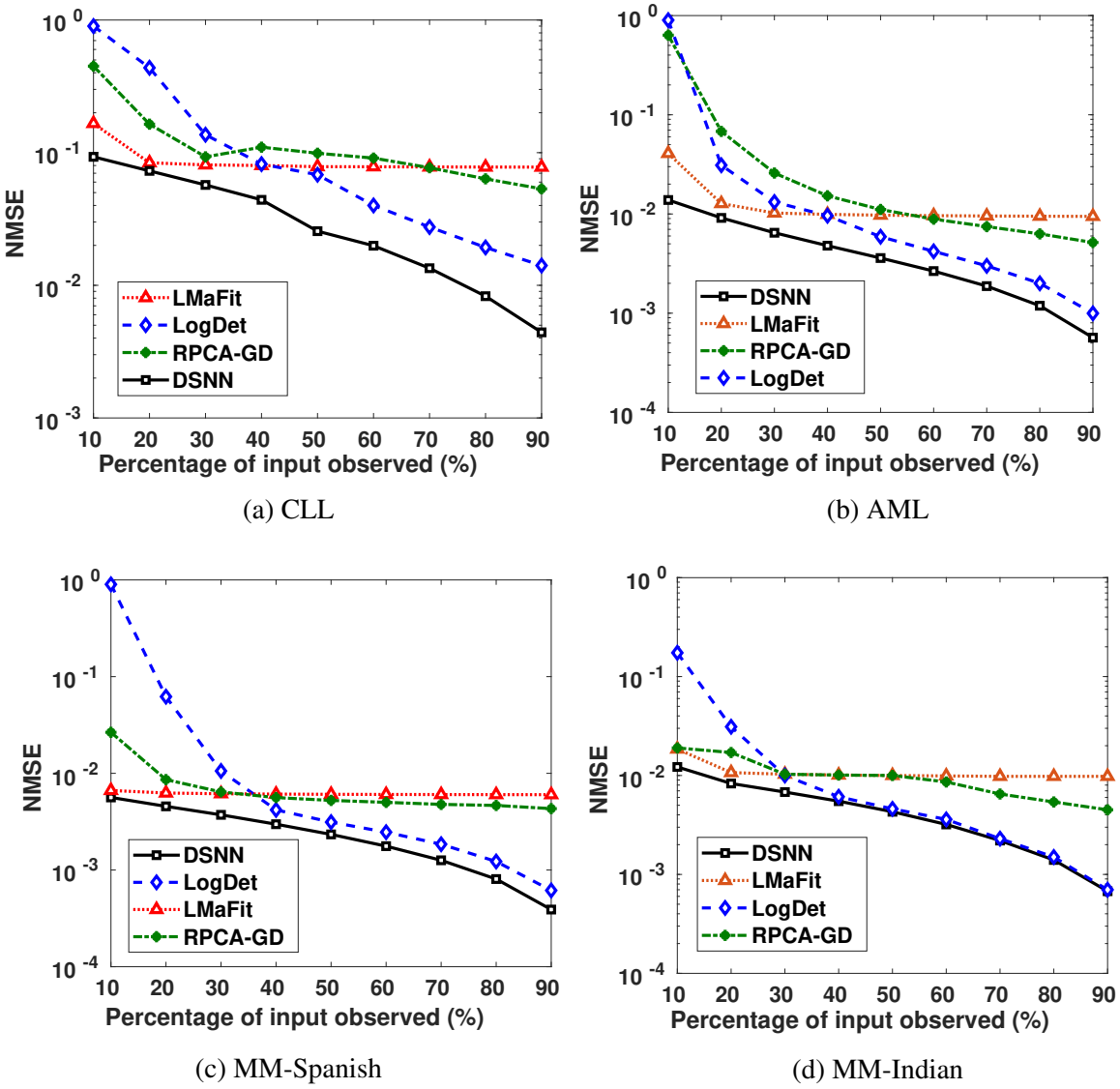


(a) CLL (b) AML

(c) MM-Spanish (d) MM-Indian

Figure 2.3: Semi-log plots with normalized y-axis show NMSE after imputation on CLL, AML, MM-Spanish and MM-Indian dataset using Stage-1 only, Stage-2 only and Proposed DSNN method (Stage-1 + Stage-2).

30 iterations. For CLL dataset, highest NMSE reported was 0.09 at 10% observed data and lowest NMSE was 0.004 at 90% observed data. For AML dataset, highest NMSE was 0.013 at 10% observed data and lowest NMSE was 0.00056 at 90% observed data. For MM-Spanish dataset, highest NMSE reported was 0.005 at 10% observed data and lowest NMSE was 0.00039 at 90% observed data. For MM-Indian dataset, highest NMSE was 0.0122 at 10% observed data and lowest was 6.25E-04 at 90% observed data.



(a) CLL

(b) AML

(c) MM-Spanish

(d) MM-Indian

Figure 2.4: Semi-log plots with normalized y-axis showing comparison of the proposed DSNN method with the three state-of-the-art methods in terms of NMSE for CLL, AML, MM-Spanish and MM-Indian dataset

2.3.2 Validation

In order to determine the significance of the DSNN method, two separate experiments were carried out on the original/true data, incomplete data and imputed data matrices. In experiment-1, classification of normal versus cancer subjects was carried out. In experiment-2, biological significance of imputation was ascertained by first identifying top candidate tumor drivers from SPARROW algorithm, followed by gene enrichment analysis on the top-ranked genes using web based application Enrichr.

2.3.3 Experiment 1: Classification

Simulation results on missing value recovery were validated by performing classification on original/true matrices, matrices with random missing values, and imputed matrices of the CLL and AML dataset. Classification can be either supervised or unsupervised depending on the availability of ground truth labels. In these dataset, ground truth labels were available. Hence, supervised classification was performed to distinguish between two classes, normal and cancer using two different classifiers: linear Support Vector Machine (SVM) and *k* nearest neighbor (KNN) method with $k = 3$. Both the dataset had large number of features, therefore, feature reduction was performed to extract important features from the data. Three different methods of feature reduction were used, Mutual Information criterion, Principal Component Analysis (PCA) and Chi-square method. Optimal number of features in each method were estimated by grid search. Further, 5-fold cross validation was performed and accuracy reported was average accuracy over 20 iterations. Experiments were performed in Python 3 environment with Sklearn 0.20 library. Classification code was written in Python programming language. Scikit-learn is a Python module for machine learning and contains various algorithms related to regression, classification and clustering. Examples of these algorithms are support vector machines (SVM), random forest (RF), *k*-means. Classification accuracy and F_1 score were calculated at different sampling ratios from 10% to 90%. The accuracy and F_1 score are defined as:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} 1(x_i = \tilde{x}_i) \tag{2.15}$$

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \tag{2.16}$$

where N is the total number of samples in the dataset, x_i is the class label of the i^{th} sample, and x_i is the class label determined by the classifier. Weighted F_1 score was used in order to account for label imbalance arising out due to unequal number of tumor and normal samples. CLL dataset had 188 tumor and 32 normal samples and AML dataset had 26 tumor samples and 38 normal. Tables 2.1 and 2.2 clearly indicate that values of classification accuracy and F_1 scores for incomplete matrices are low as compared to the values obtained on imputed matrices. Classification accuracy and F_1 scores were also computed on imputed matrices obtained from the three existing methods on both the dataset and compared with the results of DSNN method as shown in Figures 2.5 and 2.6. Classification was also performed on MM-Spanish dataset (Results are shown in Table A.1). Classification could not be performed on MM-Indian data because it was a single class data, i.e., of tumor samples only.

Table 2.1: Classification Accuracy and F_1 score for CLL dataset at varying sampling ratios (FR- Feature reduction, SR- Sampling Ratio, Obs.- Observed, Rec. - Recovered using DSNN method)

	Classification Accuracy											
FR	PCA				Chi-Square method				Mutual info method			
	KNN		Linear SVM		KNN		Linear SVM		KNN		Linear SVM	
SR	Obs.	Rec.	Obs.	Rec.	Obs.	Rec.	Obs.	Rec.	Obs.	Rec.	Obs.	Rec.
10%	.71	.87	.73	.77	.84	.96	.85	.97	.86	.96	.89	.98
20%	.71	.87	.75	.78	.80	.97	.84	.98	.86	.98	.87	.99
30%	.79	.89	.77	.81	.85	.97	.84	.98	.85	.99	.87	.99
40%	.79	.89	.81	.91	.85	.98	.85	.98	.86	.99	.88	.99
50%	.80	.89	.85	.97	.86	.99	.85	.98	.88	.99	.90	.99
60%	.78	.92	.87	.97	.85	.99	.85	.98	.90	.99	.92	.99
70%	.83	.90	.90	.97	.86	.99	.86	.98	.93	.99	.96	.99
80%	.83	.91	.96	.98	.86	.99	.87	.98	.98	.99	.99	.99
90%	.85	.91	.97	.97	.87	.98	.91	.98	.99	.99	.99	.99
	F_1 score											
FR	PCA				Chi-Square method				Mutual info method			
	KNN		Linear SVM		KNN		Linear SVM		KNN		Linear SVM	
SR	Obs.	Rec.	Obs.	Rec.	Obs.	Rec.	Obs.	Rec.	Obs.	Rec.	Obs.	Rec.
10%	.72	.86	.72	.72	.78	.96	.79	.96	.79	.95	.85	.98
20%	.72	.85	.74	.72	.77	.97	.78	.98	.79	.98	.81	.98
30%	.78	.88	.77	.77	.79	.97	.78	.97	.79	.99	.82	.99
40%	.78	.86	.80	.90	.79	.98	.79	.98	.79	.99	.85	.99
50%	.80	.88	.84	.96	.80	.99	.79	.98	.84	.99	.87	.99
60%	.78	.90	.86	.96	.79	.99	.79	.98	.88	.99	.90	.99
70%	.82	.89	.90	.97	.80	.99	.80	.98	.92	.99	.96	.99
80%	.82	.90	.96	.98	.80	.98	.82	.98	.98	.99	.99	.99
90%	.84	.91	.97	.97	.82	.98	.89	.98	.98	.99	.99	.99

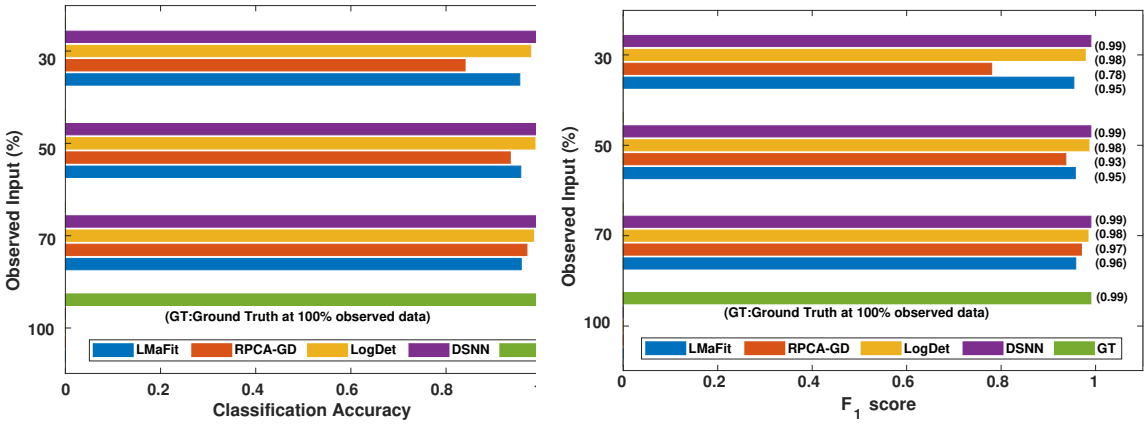


Figure 2.5: Comparison of different methods in terms of classification accuracy and F_1 score at varying sampling ratios on CLL dataset

Table 2.2: Classification Accuracy and F_1 score for AML dataset at varying sampling ratios (FR- Feature reduction, SR- Sampling Ratio, Obs.- Observed, Rec. - Recovered using DSNN method)

	Classification Accuracy											
FR	PCA				Chi-Square method				Mutual information method			
	KNN		Linear SVM		KNN		Linear SVM		KNN		Linear SVM	
SR	Obs.	Rec.	Obs.	Rec.	Obs.	Rec.	Obs.	Rec.	Obs.	Rec.	Obs.	Rec.
10%	.55	.84	.54	.83	.60	.86	.86	.96	.76	.91	.96	.98
20%	.50	.98	.50	.98	.97	.97	.98	.98	.73	.99	.91	.99
30%	.45	.99	.45	.99	.97	.97	.98	.98	.76	1.0	.91	.99
40%	.53	.99	.59	.99	.95	.99	.99	1.0	.71	1.0	.86	1.0
50%	.54	.98	.56	.99	.96	.96	.99	.99	.77	1.0	.83	.99
60%	.63	.98	.70	.99	.98	1.0	.99	1.0	.75	1.0	.93	1.0
70%	.63	.96	.67	.99	.98	.98	.99	1.0	.82	1.0	.96	.99
80%	.75	.96	.77	.99	.99	.99	.96	1.0	.87	.98	.96	1.0
90%	.80	.94	.87	.99	.99	.99	.96	.99	.94	.99	.97	.99
	F_1 score											
FR	PCA				Chi-Square method				Mutual information method			
	KNN		Linear SVM		KNN		Linear SVM		KNN		Linear SVM	
SR	Obs.	Rec.	Obs.	Rec.	Obs.	Rec.	Obs.	Rec.	Obs.	Rec.	Obs.	Rec.
10%	.53	.83	.54	.83	.48	.85	.86	.95	.76	.91	.96	.98
20%	.49	.98	.50	.98	.97	.97	.98	.98	.73	1.0	.91	.99
30%	.45	1.0	.46	.99	.97	.97	.98	.99	.76	1.0	.91	.99
40%	.52	.99	.60	1.0	.96	.99	1.0	1.0	.72	1.0	.86	1.0
50%	.53	.98	.57	.99	.96	.96	.99	.99	.78	1.0	.82	.99
60%	.64	.97	.70	.99	.98	1.0	.99	1.0	.75	1.0	.93	1.0
70%	.64	.97	.68	1.0	.98	.98	.99	1.0	.82	1.0	.96	.99
80%	.73	.96	.77	.99	.98	.99	.96	1.0	.87	.98	.96	1.0
90%	.77	.93	.87	.98	.99	.99	.96	.99	.94	.99	.97	.99

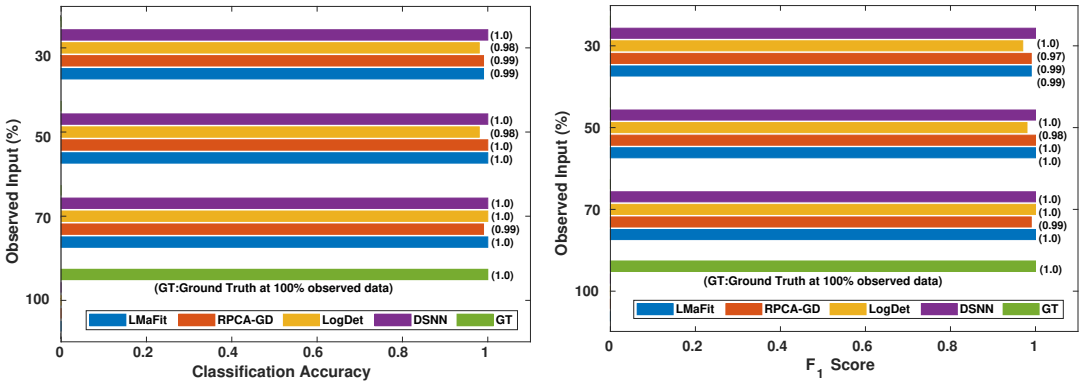


Figure 2.6: Comparison of different methods in terms of classification accuracy and F_1 score at varying sampling ratios on AML dataset

2.3.4 Experiment 2: Biological Validation

For biological validation of the results, SPARROW was applied on the original/true matrix, incomplete matrices, and imputed matrices to identify top candidate tumor driver

genes. SPARROW (SPARse selected expRessiOn regulators identified With penalized regression) was proposed by [111] and aims to find out candidate tumor drivers from the 'selected expression regulators' (SERs). It defines SERs as the genes that drive dysregulated transcription leading to carcinogenesis. In this method, variational Bayesian spike regression model has been used to fit the following model,

$$y_{m,n} = \sum x_{m,k}\beta_{k,n} + e_{m,n}, \tag{2.17}$$

where $y_{m,k}$ is the value of expression of the n^{th} gene for the m^{th} subject, $e_{m,n}$ is a normally distributed error, $x_{m,k}$ is the value of expression of the k^{th} SER for the m^{th} subject and $\beta_{k,n}$ is the additive effect of the expression of the k^{th} SER on the expression of the n^{th} gene. m ranges from 1.....M, where M is the total number of subjects and n ranges from 1.....N, where N is the total number of genes. Total SERs used in the analysis were around 3400 and they were downloaded from the link provided in the original paper. This algorithm provides a rank to each SER based on the gene expression values of the samples. The top-ranked genes from the list can be further studied by gene enrichment analysis.

For finding top 500 candidate driver genes, only the tumor samples from the data matrices were considered for SPARROW analysis. Algorithm was applied on original/true complete data matrices of all the dataset to identify the top-ranked candidate tumor drivers. This served as the ground truth for our analysis. Further, SPARROW was applied on incomplete and imputed data matrices of both the dataset at sampling ratios of 50% and 70%. Top-ranked candidate drivers from the incomplete and imputed data matrices were obtained. Gene enrichment analysis was performed on top 500 genes. KEGG pathways were studied using web based application, Enrichr, developed and maintained by [115] and [116]. KEGG pathways obtained from gene lists of original/true dataset were the ground truth. It was observed that when KEGG pathway analysis was done for incomplete matrices, these were not able to predict cancer pathways with a higher significance (low *p*-value) whereas for imputed matrices, cancer pathways were predicted with a higher significance due to decrease in *p*-value. Results from KEGG analysis on all dataset are presented in tabular form showing the *p*-values, combined score for original/true data and the incomplete and complete matrices in tabular form in the Tables A.2, A.3, A.4 and A. *p*-value was computed from the Fisher exact test. Fisher test was run on random gene sets and ranks were derived at each run. Mean rank was calculated from the different runs and standard deviation of the rank obtained from the expected rank was also calculated for each term in the gene-set library. Finally, a z-score was calculated to estimate the deviation from the expected rank. z-score and *p*-value were used to compute combined score which is obtained by multiplying z-score with the logarithm of *p*-value. A detailed analysis for CLL dataset consisting of z-score

and combined score has also been shown in the Tables A.6 and A.7.

2.4 Discussion

2.4.1 Importance of the proposed DSNN method

DSNN, a two stage method proposed for matrix recovery was based on Compressive Sensing Framework. In Stage-1, it utilized column and row sparsity of the gene expression matrix in DCT domain for missing value imputation, while in Stage-2, it exploited low rank nature of the matrix for denoising. Expression values of any particular gene would vary slowly across subjects, thereby, exhibiting sparsity in columns in some transformed domain. Similarly, expression values of a subject for most of the genes will also be slowly varying, thereby, exhibiting sparsity in the rows. Since there is a high inter-dependence between the expression levels of the genes, one may consider gene expression matrix as a low rank matrix. Thus, as discussed earlier, both the assumptions used in Stage-1 (of sparsity in DCT domain) and Stage-2 (low rank of matrix) hold true for the given gene expression data. This work utilizes double sparsity, i.e., sparsity on both the columns and the rows in the DCT domain. Most of the imputation algorithms developed for missing value imputation such as KNN, LSimpute, LLSimpute, BPCA etc. work at high observability of data, while the proposed DSNN method worked well even when data had very high missing rates of 10% to 40%. The proposed DSNN method performed better than the other matrix completion methods at all sampling ratios. The state-of-the-art matrix imputation methods that have been used for performance comparison in this work required a lot of parameter tuning for optimal performance, while DSNN method did not require parameter tuning to such a great extent.

2.4.2 Improvement in Classification Accuracy

It was evident from the results shown in Tables 2.1 and 2.2 that the classification accuracy and F_1 scores reduced as the number of missing values increased. There were 220 samples in CLL dataset and 64 samples in AML dataset. For smaller dataset like AML, missing values affected the classification accuracy and F_1 scores greatly. Thus, it is necessary to impute missing values in gene expression data to prevent incorrect downstream analysis of the data. When the classification was performed on the imputed data, there was considerable improvement in the classification accuracy, thereby, validating our hypothesis. Classification accuracy and F_1 scores calculated on original/true com-

plete data matrices (100% sampling ratio) were considered as ground truth values. For CLL dataset, ground truth values of classification accuracy and weighted F_1 score were 0.99 and 0.99, respectively, as shown in Figure 2.5. For KNN classifier and Chi-square feature selection approach, classification accuracy and F_1 score obtained for 50% observed data was 0.86 and 0.80 respectively as shown in Table 2.1. After imputation, values improved significantly to 0.99 and 0.99. For AML dataset, ground truth values of classification accuracy and F_1 score were 1.0 and 1.0 respectively as shown in Figure 2.6. Similarly for Linear SVM classifier and PCA feature selection approach, classification accuracy and F_1 score for 50% observed data was 0.56 and 0.57, respectively, as shown in Table 2.2. After matrix imputation, classification accuracy and F_1 score improved considerably to 0.99 and 0.99 respectively. For every sampling ratio, consistent results were obtained that validates our method.

Improvement in functional enrichment analysis for KEGG pathways

KEGG enrichment analysis was performed on the top 500 ranked genes obtained from SPARROW algorithm to biologically validate our results. As mentioned earlier, KEGG pathways obtained by the top-ranked genes of original/true matrices were considered the ground truth values. Pathways with *p*-value < 0.05 were only considered. When KEGG analysis was done on top-ranked genes from incomplete matrices, there was significant decrease in the *p*-value of the most significant pathways. “Wnt signaling pathway" [127, 128] and “Notch signaling pathway" [129, 130] are important pathways in CLL cancer. An important observation was that *p*-value for “Notch signaling pathway" was 2.00E-01 at ground truth and it was 5.76E-02 at 70% observed data for CLL dataset. Values were insignificant in both the cases. However, after imputation, *p*-value became significant with value 1.56E-02 which was less than 0.05 as shown in Figure 2.7.

Similarly, *p*-value for “Wnt signaling pathway" was 8.33E-05 on original/true dataset, as shown in the Figure A.1. At 50% observed data p-value for “Wnt signaling pathway" was 3.10E-02 which was less significant than the ground truth value at 50% observed data. After matrix imputation, *p*-value became significant with value 2.13E-03. Similarly, *p*-value became 1.90E-05 after matrix imputation on 70% observed data which was more significant than the *p*-value 6.66E-5, observed at 70% data. “Fc epsilon RI signaling pathway" is an important pathway in AML cancer [131]. This pathway was insignificant for original/true data with *p*-value 2.12E-01. At 70% observed data, *p*-value was 9.40E-02 which was again greater than 0.05. After matrix imputation, the value became significant at 2.75E-02, which was less than 0.05 as shown in Fig. 2.8. Similarly, ‘Ras signaling pathway" is activated in Multiple Myeloma cancer [132]. For MM-Spanish data, “Ras signaling pathway" was significant with *p*-value 0.0052 for

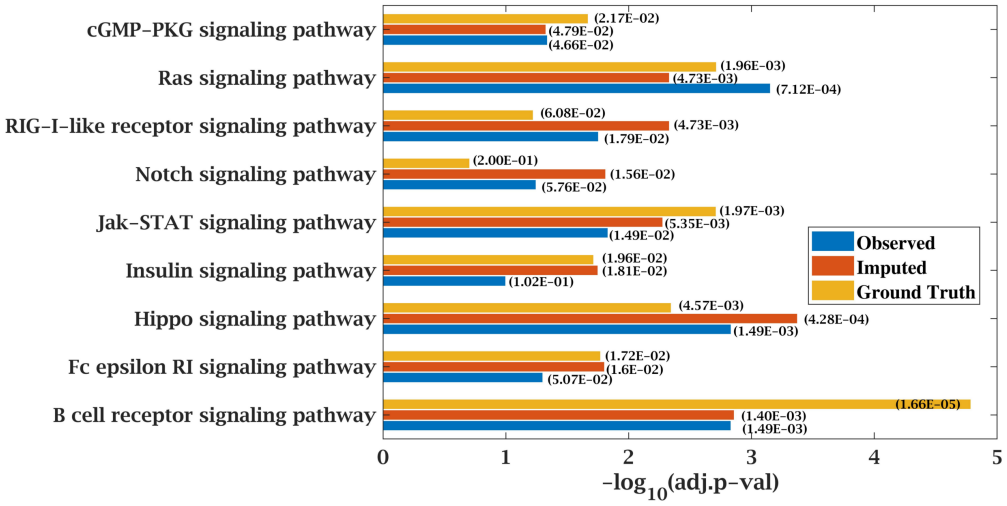


Figure 2.7: Few important KEGG pathways at 70% observed and imputed data for CLL data. Adjusted p-values are shown in brackets.

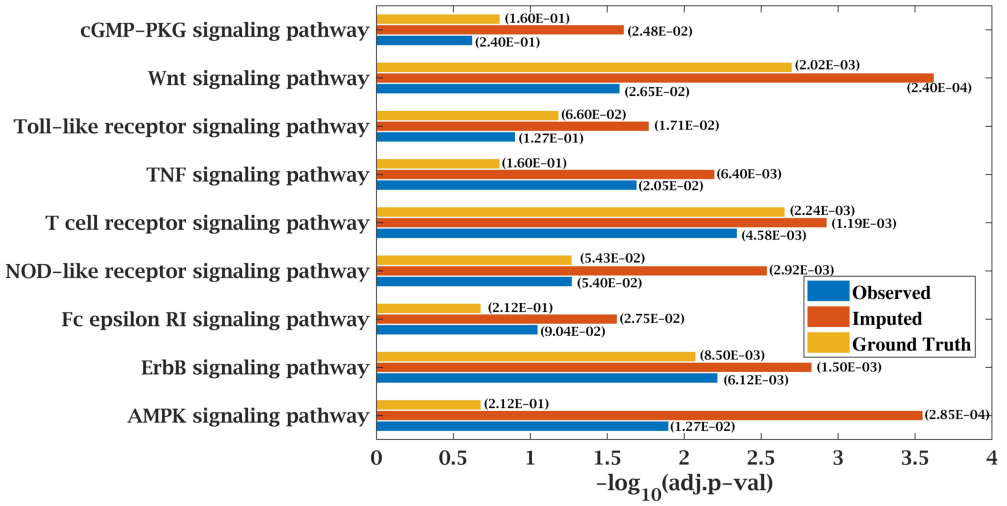


Figure 2.8: Few important KEGG pathways at 70% observed and imputed data for AML data. Adjusted p-values are shown in brackets.

original/true data but became insignificant with p-value 0.23 when 70% data was observed as shown in Figure 2.9. After matrix imputation, significance of the pathway was restored with p-value 0.04. For MM-Indian dataset,"Transcriptional misregulation in cancer" was found to be insignificant with p-value 0.55 as shown in Figure 2.10. After imputation, p-value decreased to 1.37E-03 and became more significant than ground truth p-value, 7.8E-03. Additional KEGG analysis results on the dataset CLL, AML and MM-Spanish data are provided in the Figures A.2, A.3 and A.4.

Thus, DSNN method not only imputed missing entries but also performed some denois-

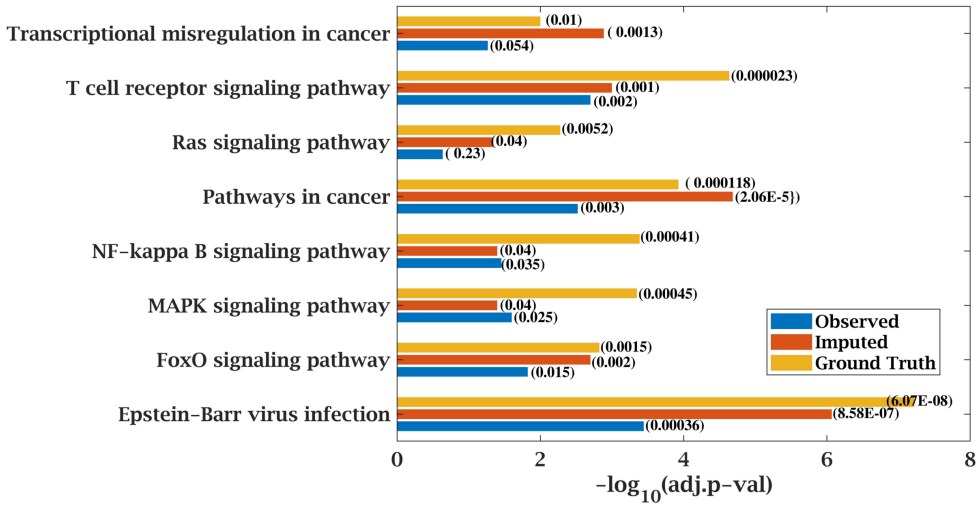


Figure 2.9: Few important KEGG pathways at 70% observed and imputed data for MM-Spanish data. Adjusted p-values are shown in brackets.

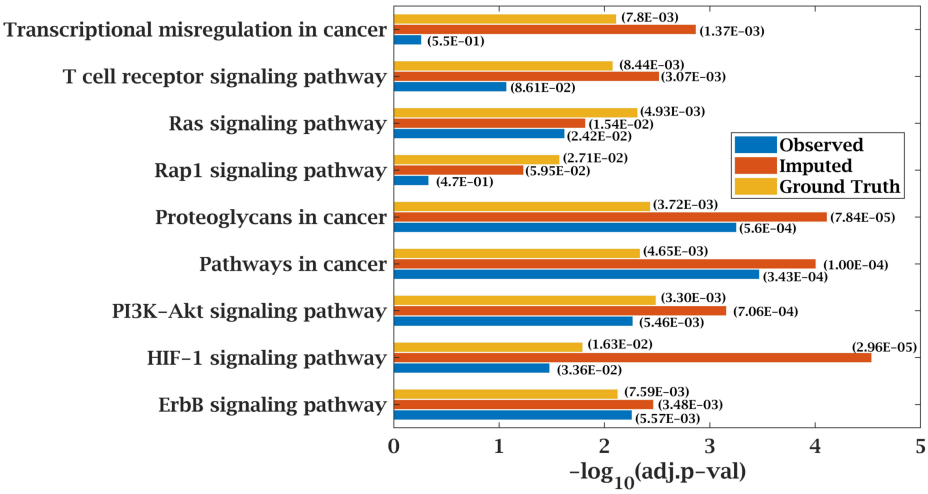


Figure 2.10: Few important KEGG pathways at 70% observed and imputed data for MM-Indian data. Adjusted p-values are shown in brackets.

ing to improve the results. It is quite evident from the analysis that gene enrichment analysis results were partially inaccurate due to incomplete matrices. This was because the genes identified as top-ranked genes by performing SPARROW analysis on complete data matrix were not identified in the top-ranked list obtained from incomplete data matrix. However, when the incomplete matrix was imputed using the proposed DSNN method, top-ranked list of genes obtained from SPARROW analysis was quite similar to the ground truth. Our observation demonstrates the importance of imputing missing values in gene expression data.

2.5 Conclusion

Gene expression data generally has a lot of missing values that can adversely influence the downstream analysis. Hence, missing value imputation in gene expression data is important for appropriate analysis in cancer research. In this work, we have proposed an optimization based method for imputing missing values in the gene expression data using discrete cosine transform based sparsity and nuclear norm minimization. The proposed method is validated quantitatively based on the application of classification. Additionally, we have also biologically validated the significance of imputation by performing pathway enrichment analysis. The proposed method is tested on datasets of hematological malignancies involving CLL, AML and MM. While working with multiple myeloma (MM) dataset, we got interested in pursuing research in MM. MM is a type of blood cancer where there is presence of abnormal plasma cells in the blood. One of the main challenges in MM is monitoring disease progression and dealing with drug resistance in patients which often leads to poor outcome. Therefore, longitudinal studies involving MM patients might help in identifying the genomic events leading to disease progression and drug resistance. Given the significance of the mentioned research problem, we decided to study MM. The study has been presented in detail in next chapter.

Chapter 3

Clonal evolution in Multiple Myeloma

3.1 Introduction

Multiple Myeloma is a hematological malignancy characterized by clonal expansion of abnormal plasma cells in the bone marrow. Patients with MM show symptoms of calcium elevation, renal failure, anemia and bone lesions as defined in CRAB criteria. Due to heterogeneity in MM, survival outcomes may vary and needs continuous monitoring of patients as drug resistance and disease progression are widely seen in MM. Hence, as discussed in section 1.2, longitudinal data of MM needs to be studied to gain deeper understanding of the genomic alterations taking place in MM as the disease evolves. In this study, we have evaluated 186 pairwise whole exome sequences obtained from 62 MM patients at two time points representing tumor at diagnosis, tumor at progression and compared to their germline landscapes respectively using NGS. We have identified individual clonal genomic complexities, tumor mutation burdens (TMBs) and divergence of clusters of mutations in founder clones. This study has provided novel insights into recurrent subclonal shifts in drivers (DRV), oncogenes (ONC), tumor suppressor genes (TSGs) and the potential actionable targets (ACT) associated with progression of MM.

3.2 Materials and Methods

This study was approved by the Institute Ethics Committee and conducted as per ethical guidelines. Voluntary written informed consent was obtained from all the study individuals.

3.2.1 Whole exome sequencing

Genomic DNA was isolated from CD138+ plasma cells enriched from bone marrow aspirates with MACS magnetic microbeads (Miltenyi Biotech, Germany), collected from 62 patients including 61 newly diagnosed treatment naïve MM patients and 1 MGUS (who later converted to MM at TP2) diagnosed as per IMWG guidelines (Table 3.1). Patients diagnosed and treated at our center from 2014 to 2019 in whom DNA samples were available prior to therapy and at the time of disease progression were included in

this study. The patients were treated with triplet combination induction chemotherapy - VCD (bortezomib, cyclophosphamide, dexamethasone) or VTD (bortezomib, thalidomide, dexamethasone) or VRD (bortezomib, lenalidomide, dexamethasone) prior to time of progression. The median overall survival (OS) of the patient cohort was 152.5 weeks and median progression free survival (PFS) was 87.21 weeks.

Whole exome sequencing (WES) was carried out on 186 DNA samples extracted from 62 MM patients collected at two time points- one prior to any therapy at diagnosis (Time Point 1= TP1) and second at a follow up time point of disease progression (Time Point 2= TP2). WES was also carried out on paired germline DNA obtained from peripheral blood mononuclear cells for all the patients.

For WES, DNA was extracted using Maxwell RSC cultured cells DNA kit (Promega, Wisconsin, USA) on automated nucleic acid extraction system (Promega, Wisconsin, USA). Prior to library construction, DNA was quantified fluorometrically with a DNA high sensitivity kit with Qubit (ThermoFisher Scientific, MA, USA). WES libraries were constructed from genomic DNA using the Nextera Exome kit (Illumina, San Diego, California, USA) which targets a genomic footprint of 62Mb with >3,40,000, 95 mer probes. After quantification, the DNA was normalized to 10ng/µl and a total of 50ng DNA was tagmented with transposons. The tagmented DNA was purified from the transposome with sample purification beads. The purified tagmented DNA was subjected to a unique combination of dual index adapters and amplified with sequences required for cluster generation. After amplification, the DNA libraries were purified and the purified libraries containing unique indices were combined into a single pool using a 3-plex strategy. The target regions of interest in the purified libraries were hybridized with coding exome oligos and captured with streptavidin magnetic beads. The enriched libraries were eluted from the beads and subjected to a second round of hybridization with coding exome oligos. Final libraries were eluted and then quantified and evaluated for quality using DNA high sensitivity Qubit kit (ThermoFisher Scientific, MA, USA) and DNA HS Kit (Agilent Technologies, Santa Clara, USA) on Agilent Bioanalyser respectively. The size range of generated libraries was 200-500 bp. The resultant captured libraries were pooled, normalized following standard normalization method and paired-end sequencing was carried out using the Illumina cBot system and HiSeq SBS kit V4-250 cycles on HiSeq 2500 (Illumina).

3.2.2 Analysis of Whole exome data

The overall workflow of data analysis is shown in Figure 3.1. Raw sequencing reads were quality checked using FastQC software (v0.11.4, `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`). The adapter sequences were removed

Table 3.1: Baseline demographic, laboratory and clinical characteristics of multiple myeloma (MM) patients (n = 62)

Parameter	**No. of patients**
Median Age (Range) In Years	58 (31 to 72)
Gender	
Male	38
Female	24
Hemoglobin (g/dL)	
≤ 10	39
> 10	23
Platelet Count (/dL)	
< 100	10
≥ 100	52
Serum creatinine (mg/dL)	
$\leq= 2$	49
> 2	13
Serum albumin (g/dL)	
< 3.5	30
≥ 3.5	32
ISS 1 / 2 / 3	1/17/44
RISS I / II / III / NA	1/36/14/11
MRS 1 / 2 / 3 / NA	7/33/21/1
Serum calcium (mg/dL)	
0-11	54
> 11	8
eGFR (mL/min)	
< 40	17
≥ 40	45
IgG Isotype	
IgA	14
IgG	37
Light chain κ/λ	11
BM plasma cells (%)	
≤ 40	21
> 40	41
Serum LDH (IU/L)	
$\leq$420	51
> 420	6
NA	5
β2-microglobulin (mg/L)	
< 3.5	3
$\geq$3.5	59

using Trimmomatic software (v0.39, http://www.usadellab.org/cms/?page=trimmomatic). Illumina Dragen somatic pipeline (v3.5.7, https://sapac.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/edico-genome-inc-dragen-somatic-pipeline.html) was used to process the trimmed reads and aligned with human reference genome, hg19 available at UCSC.

The tumor and normal bam files obtained from Illumina Dragen somatic pipeline were used for variant calling using three additional variant callers, Strelka v2.9.10 [133]; SomaticSniper v1.0.5.0 [134] and SpeedSeq v0.1.2 (FreeBayes)[135] in order to validate the variants called by Dragen somatic pipeline. Only those variants called by all the four callers and passed filters of base quality ($\geq$ 20), mapping quality ($\geq$ 20), tumor reads ($\geq$ 10) and normal reads ($\geq$ 5) qualified as a consensus. These validated variants were further annotated using BaseSpace Variant Interpreter (https://variantinterpreter.informatics.illumina.com/home).

Further, COSMIC database was explored for assignment of variant pathogenicity (Pathogenic / Neutral / Unknown). Variants predicted as Deleterious / Damaging / Pathogenic by any of the three tools (SIFT / PolyPhen / FATHMM) were considered as Pathogenic. For identification of CNVs, the .bam files of tumor and normal samples obtained from Illumina Dragen (v3.3.7) somatic pipeline were analyzed using Sequenza (https://cran.r-project.org/web/packages/sequenza/) package along with human reference .fasta file from UCSC (ucsc.hg19.fasta).

Variants identified were compared with the variants identified in MMRF CoMMPass Study database (www.themmrf.org). The mutated genes were classified as driver genes, oncogenes and tumor suppressor genes based on publicly available resources listed at cBioPortal [136, 137] (https://www.cbioportal.org/); at intOgen (https://www.intogen.org/search) ([138]; OncoKB (https://www.oncokb.org/) [35] and as described by 2014 [72].

Potentially actionable targets were identified in this study based on repository of FDA approved on label or off-label drugs or those experimentally druggable compiled and listed in literature [139, 140], at the TARGET (Tumor Alterations Relevant for Genomics driven Therapy) (https://software.broadinstitute.org/cancer/cga/target) database of the Broad Institute and the COSMIC actionability data v93 (https://cancer.sanger.ac.uk/cosmic). The TARGET database is a database of genes that when somatically altered in cancer, are directly linked to a clinical action. The tumor mutational burden (TMB) defined as the number of nonsynonymous mutations/ Mb was calculated from average coverage with respect to total bases (3137161264) in binary mode and with reference to human genome (hg19). Clonal evolution patterns were evaluated using QuantumClone (https://www.rdocumentation.

`org/packages/QuantumClone/versions/0.15.11`) [141] and the cellular prevalence values, $\hat{\theta}$, were calculated as defined below.

$$\hat{\theta} = VAF \times \frac{N_{Ch} + N_{Ch(Normal)} \times \frac{1-p}{p}}{NC} \tag{3.1}$$

where VAF stands for variant allele frequency, N_{ch} is the number of copies of the corresponding locus in cancer cells, $N_{Ch(Normal)}$ is the number of copies of the corresponding locus in the normal cells ($N_{Ch(Normal)}$ = 2 for autosomes) and NC is the number of chromosomal copies bearing the variant and p is the tumor purity. VAF is the ratio of the number of reads supporting variants/mutations divided by the total number of reads at the particular position [142, 141]. NC is a priori unknown and is deduced by the QuantumClone [141].

The cellular prevalence values $\hat{\theta}$ of each cluster obtained from QuantumClone were subjected to fishplot R package for visualization[143]. Cellular prevalence values higher than 1 were set to 1 as suggested[141]. Clonal patterns were classified as branching or linear or stable as described[68]. In case of branching evolution, both gain and loss of clones was observed. In case of linear evolution, there was gain of mutations but no clonal loss; while in stable progression, the clonal structure remained preserved at two time points. Stable with loss pattern had predominantly conserved clonal structure but there was also evidence of clonal loss at a subsequent time point. The biological pathways relating to altered clonal mutational profiles were deduced by gene enrichment analysis using Enrichr (https://maayanlab.cloud/Enrichr/) as described [116].

3.2.3 Statistical Analysis

Clinical and biological characteristics of the patients were analysed using Chi-squared or Fisher's exact test for discrete categorical variables as applicable. Nonparametric statistical analysis was carried out for continuous variables with Wilcoxon signed rank test. A p-value of <0.05 was considered statistically significant.

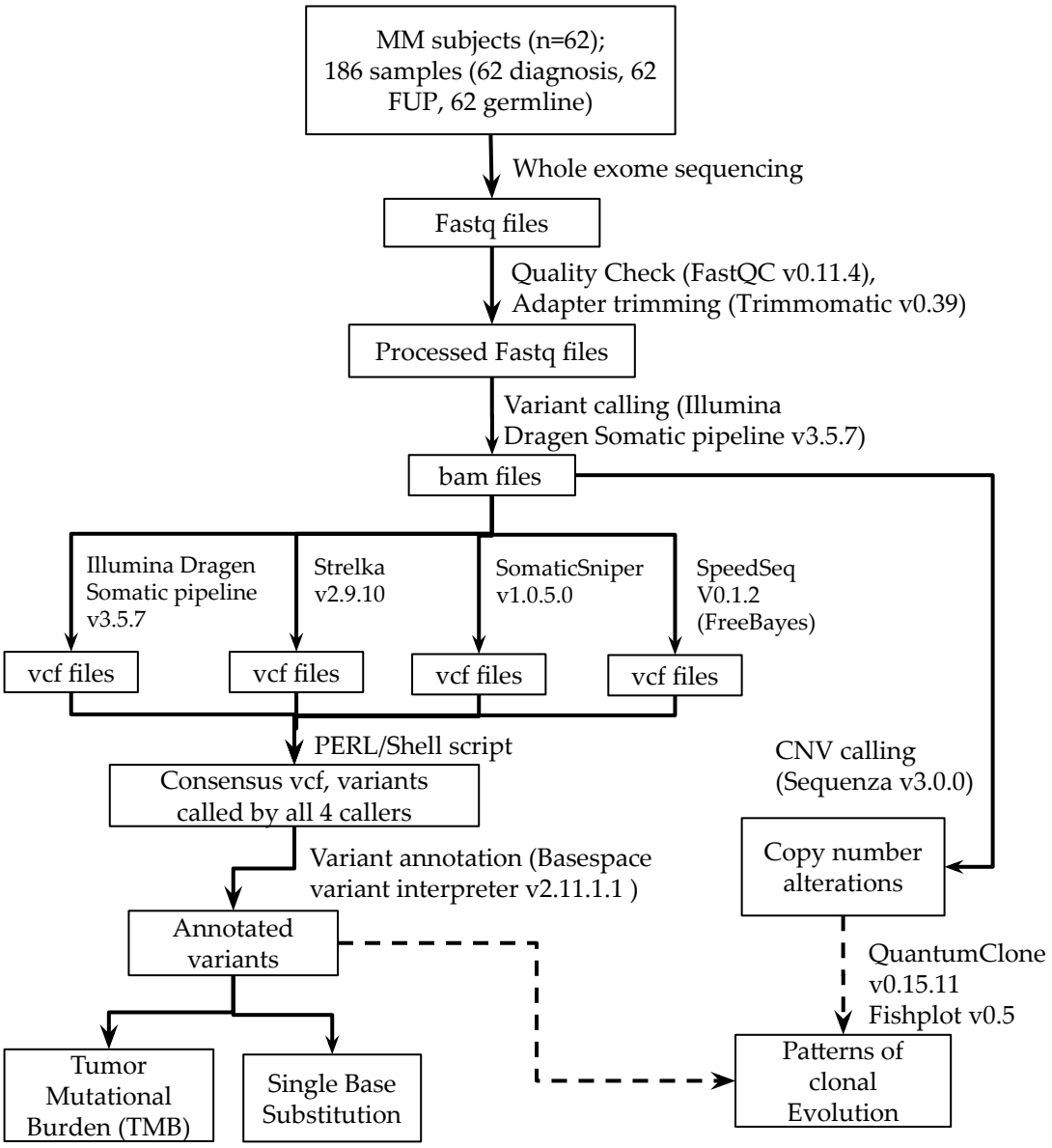


Figure 3.1: Workflow of Study and data analysis. Analysis workflow of the WES study performed on 62 MM patients whose tumor PC samples were sequenced at diagnosis, at follow up and compared with their germline profiles. Fastq files were quality checked with FastQC, adaptors trimmed with Trimmomatic and processed further through Illumina Dragen Somatic pipeline for variant calling. Variants were validated with additional 3 variant callers (Strelka2, SomaticSniper and SpeedSeq), a consensus .vcf was derived and annotated with Variant Interpreter for deducing TMB and SBS with Sigprofiler. CNVs were identified with Sequenza and processed further with QuantumClone and Fishplot for interpretation of patterns of clonal evolution.

3.3 Results

3.3.1 Estimation of somatic mutations at two time points

A total of 13951 and 11684 nonsynonymous (NS) somatic mutations were identified in myeloma pairwise whole exomes sequenced at diagnosis (TP1) and at progression (TP2) respectively (Table 3.2). Among these, 4410 somatic mutations in TP1 and 3833 in TP2 were classifiable as pathogenic. At diagnosis, 10561 somatic mutations were missense type, 1227 belonged to 3′, 1437 were in splicing sites and 538 mapped in 5′UTR regions. On progression, these reduced to 8996, 946, 1207 and 375 somatic mutations representing missense, 3′, splicing and 5′UTR mutations, respectively. The average numbers of somatic mutations/ sample at diagnosis totalled 236.45 at TP1 while 198.03 at TP2 (Table 3.2). At TP1, there were an average of 179 missense mutations/sample (152.47 at TP2), followed by 20.8 in 3′ (16.03 at TP2), 24.36 in splicing regions (20.46 at TP2), and 9.12 in 5′UTR region (6.36 at TP2). Patients with high somatic mutations may possess high neoantigen loads and may benefit from immunotherapies.

3.3.2 Tumor mutation burden declines from diagnosis to progression in hypermutators

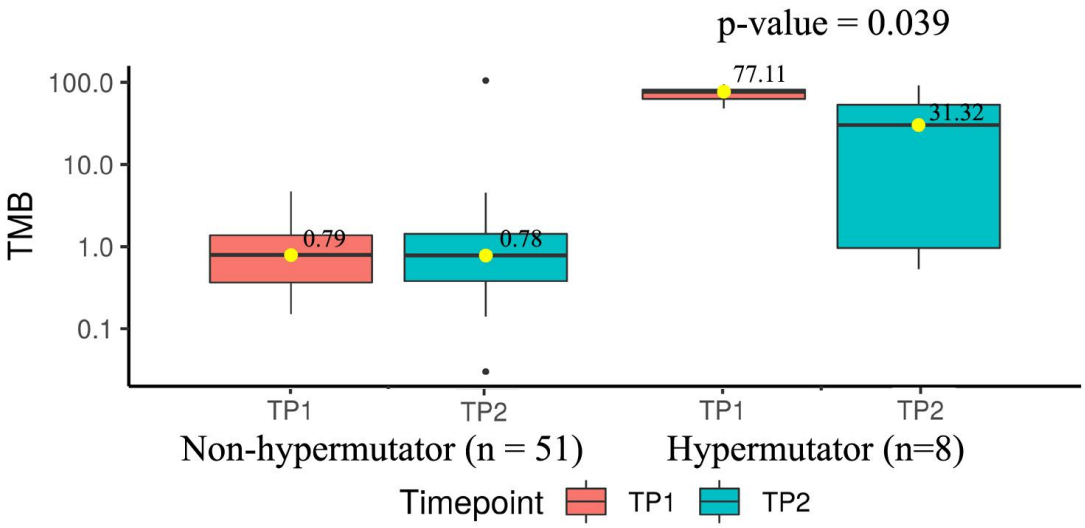


Figure 3.2: Changes in TMB at diagnosis and on progression Comparison of median TMB across MM patients at TP1 and TP2 in non-hypermutator (n=51) (TMB<10) and hypermutator category (n=8) (TMB between 10 to 100)

Table 3.2: A comparison of number of nonsynonymous (NS) somatic mutations, tumor mutation burden (TMB) and single base substitutions (SBS) in MM at diagnosis and on progression

Type of somatic mutations	**Time point**	
	TP1	**TP2**
IN ALL SAMPLES (n=59)		
Number of somatic mutations	13951	11684
Number of known pathogenic somatic mutations	4410	3833
Number of Missense somatic mutations	10561	8996
Number of Nonsense somatic mutations	188	160
Number of somatic mutations in 3'UTR	1227	946
Number of somatic mutations at Splicing sites	1437	1207
Number of somatic mutations in 5'UTR	538	375
MEANS PER SAMPLE		
Average number of somatic mutations/sample	236.45	198.03
Average number of Missense somatic mutations/sample	179	152.47
Average number of Nonsense somatic mutations/sample	3.19	2.71
Average number of somatic mutations in 3'UTR/sample	20.8	16.03
Average number of somatic mutations at Splicing sites/sample	24.36	20.46
Average number of somatic mutations in 5'UTR/sample	9.12	6.36
MEDIAN number of NS somatic mutations	32	34
Tumor Mutation Burden (TMB)		
MEDIAN TMB	0.85	0.93
AVERAGE SBS IN ALL SAMPLES		
C>T	128.88	101.86
T>C	85.02	64.66
C>A	34.34	27.56
C>G	28.64	21.92
T>G	21.98	16.47

Patients at diagnosis had an average tumor mutation burden (TMB) of 10.8 NS somatic mutations/Mb/sample (range 0.15 to 95) that reduced to 7.46 (range 0.03 to 105.47) on progression. The median TMB among patients at TP1 and TP2 were 0.85 and 0.93 respectively. The median TMB at two time points among patients with age at diagnosis $\leq$ 65 years (0.82 versus 0.76) and those with >65 years (1.62 versus 1.22) were comparable.

Patients were classified on the basis of their TMB levels at diagnosis as those with low TMB of $\leq$ 10 (n= 51) and high TMB levels $\geq$ 10 to $\leq$ 100 (n=8) (i.e., hypermutators). Three patients (SM0007, SM0052 and SM0145) were outliers or super-hypermutators with $\geq$ 100 TMBs (134.43, 132.12 and 126.3 respectively) and were analyzed for clonal evolution exclusively. In particular, patients grouped into high TMB category (TMB levels $\geq$ 10 to $\leq$ 100) had median TMB levels at TP1 (77.11) that significantly reduced

Table 3.3: Classification of genes harbouring NS somatic mutations and the variants observed in MM in this study

Classification	**Number of genes with mutations (n=8977)**	**Number of mutations (n=19022)**
Known to be mutated in some cancer	8869	18817
Known to be mutated in MM	7107	15864
Mutated in MMRF CoMMPass study	6690	15063
Known oncogenes	131	252
Known tumor suppressor genes	176	443
Known to be driver genes in some cancer	320	821
Known to be driver genes in MM	72	221
Known as actionable (COSMIC)	100	239
Drivers with decreased frequencies on progression	39	140
Drivers with increased frequencies on progression	12	36
Drivers with constant frequencies both at diagnosis and on progression	21	45

at TP2 (31.32; p=0.039) (Figure 3.2). Hypermutators might sustain stable drug resistant clones and hence may benefit from combinations of IMiDs with novel therapeutics.

3.3.3 Comparison of frequencies of driver genes mutated at diagnosis versus progression

Table 3.3 summarizes number of mutated genes and mutations that were encountered in MM in this study. Out of 8977 total mutated genes that got shortlisted, 8869 were found to be mutated in some form of cancer while 7107 genes were identified to be mutated in MM among which 6690 genes have been reported in MMRF CoMMPass dataset. A set of 131 mutated genes turned out to be known oncogenes, 176 were established tumor suppressors, 320 were known drivers across different cancers while 72 genes were found to be known driver genes in the context of MM. Of all these genes harbouring somatic mutations in MM, 100 genes got classified as COSMIC candidate actionable targets.

We screened the WES data for a total repertoire of 102 known driver genes for MM and found 72 driver genes to be mutated. We then analyzed which driver genes had subclonal gains or losses or remained stable with progression and arranged them in descending and ascending series (Figure 3.3). These drivers were further shortlisted to those that had topmost number of recurrent subclonal shifts and were observed in atleast 3 or more patients. Figure 3.3a shows topmost temporal falls in PABCP1, BRAF,

KRAS, CR1, DIS3, ATM and other genes while Figure 3.3c shows topmost temporal increases that were observed in KMT2C, FOXD4L1, SP140 and NRAS. Similarly, Figure 3.3b shows the most recurrent drivers like FAT4 and IGLL5 that remained stable on progression. Contrasting mutational landscapes at diagnosis and at progression highlight the importance of their immediate monitoring prior to tailoring therapy.

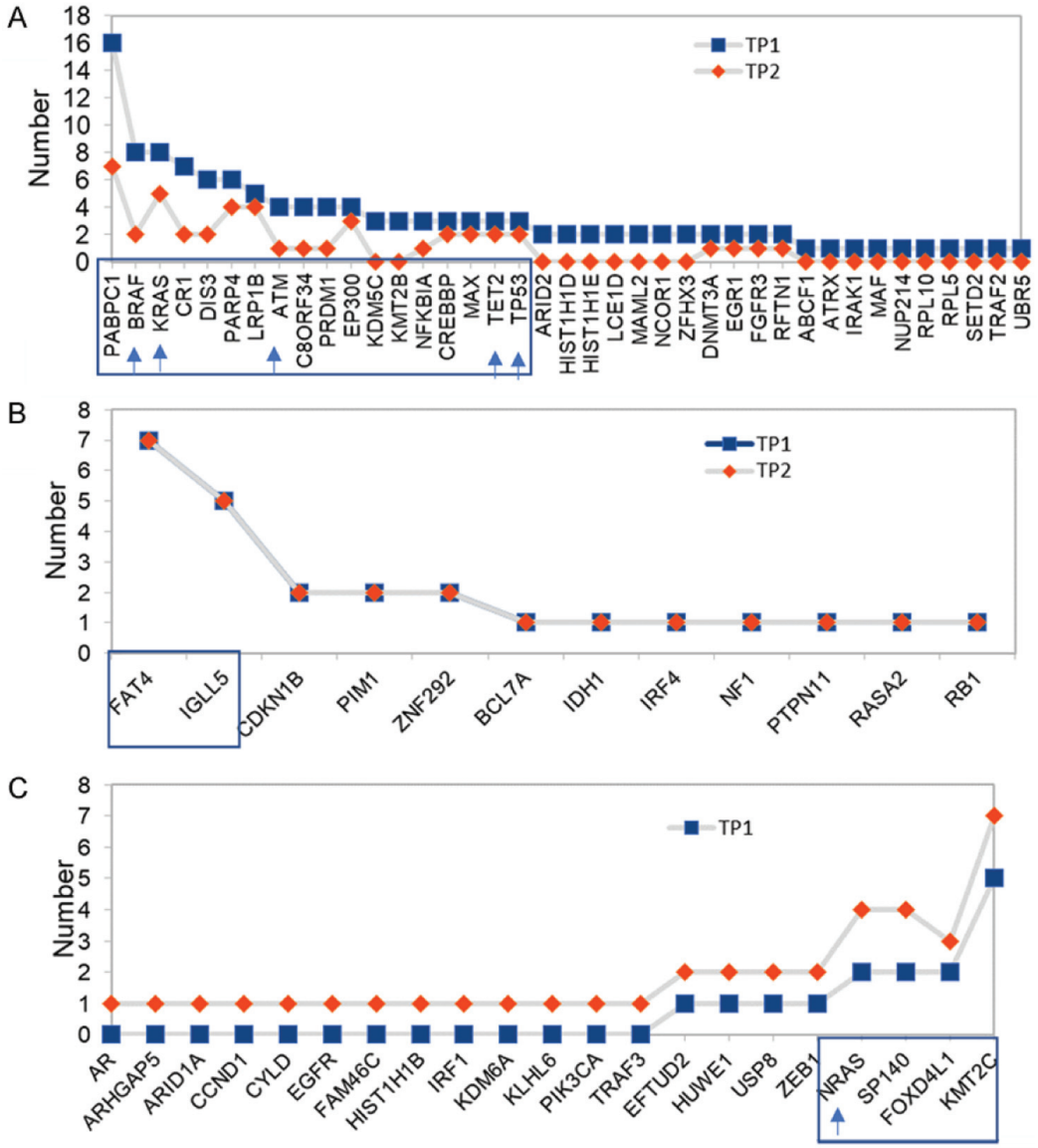


Figure 3.3: Temporal changes in distribution of driver genes on progression. Distribution of mutated driver genes in MM patients at TP1 and compared to TP2. (A) Falling mutated drivers whose frequencies decreased in TP2, (B) Drivers that are maintained at constant frequencies throughout the disease, and (C) Rising mutated drivers whose preponderance increased in patients at TP2. Driver mutation profiles observed in atleast 3 or more patients are shown inside boxed frames. Actionable genes are indicated by arrows on X axis.

3.3.4 Distribution of mutated potential actionable target genes at diagnosis and progression

As many as 19022 somatic mutations (Table 3.3) were observed at varying frequencies among 8977 genes in MM patients in this study. Of these, 18817 variants are known mutants in cancers of some kind, 15864 have been reported to be mutated in MM while 15063 have been described in MMRF dataset. These consisted of 821 mutations across drivers known to be associated with different cancers and 221 mutations in 72 driver genes (*BRAF, SP140, EP300, FAT4, PABPC1, CREBBP, FOXD4L1, PRDM1, KMT2C, C8ORF34, NRAS, KRAS, DIS3, NFKBIA, LRP1B, IGLL5, ZNF292, ATM, CR1, PTPN11, BCL7A, CDKN1B, PARP4, RB1, MAX, NF1, EFTUD2, TP53, DNMT3A, RASA2, RFTN1, TET2, EGR1, HIST1H1E, PIM1, ZEB1, FAM46C, LCE1D, CCND1, MAML2, ARID2, ARID1A, TRAF3, ARHGAP5, USP8, CYLD, ZFHX3, MAF, NCOR1, RPL5, KMT2B, IDH1, PIK3CA, KLHL6, SETD2, FGFR3, IRF1, HIST1H1D, HIST1H1B, ABCF1, IRF4, EGFR, UBR5, NUP214, TRAF2, IRAK1, RPL10, KDM6A, KDM5C, HUWE1, AR, ATRX*) known to be involved in MM. There were 252 somatic mutations in oncogenes, 443 in tumor suppressor genes and finally 239 variants were found across 100 potential actionable genes.

Table 3.4 summarizes a list of variations in 22 actionable target genes that were found mutated in atleast 3 patients at either or both time points. These consisted of *BRAF, FANCM, MRE11, WRN, EXO1, FANCA, ALK, FANCD2, MSH3, NBN, NRAS, KRAS, FLT3 MAP2K1, PALB2, RAD51D, RAD51C, MERTK, KDR, RAD54B, FANCG, PTCH1*. The most common actionable mutation was Val600Glu in *BRAF* that was most abundant at the time of diagnosis. Identification of druggable targets at subclonal levels could aid in treating patients with genome defined target specific drugs.

3.3.5 Comparison of Single nucleotide substitutions at diagnosis and progression

As shown in Table 3.2, six types of single base substitutions (SBS) were observed. The SBS C>T was the most predominant form of mutation found both at TP1 (128.88; 40.94%) and TP2 (101.86; 41.63%) followed by T>C (85.02; 27.017% at TP1, 64.66; 26.42% at TP2), C>A (34.34; 10.9% at TP1, 27.56; 11.26% at TP2), C>G (28.64, 9.09% at TP1; 21.92, 8.95% at TP2), T>G (21.98, 6.98% at TP1; 16.47, 6.73% at TP2), T>A (15.9, 5.05% at TP1; 12.25, 5% at TP2).

Table 3.4: Frequency of variations in actionable genes observed in atleast 3 or more multiple myeloma patients

VARIANT	REF	ALT	EXON	HGVSC	HGVSP	CONSEQUENCE	GENE	Number of patients with mutation	Count (TP1)	Count (TP2)
7:140453136:T	A	T	15/18	c.1799T>A	p.(Val600Glu)	missense_variant	BRAF	7	5	2
14:45606287:T	C	T	2/23	c.524C>T	p.(Ser175Phe)	missense_variant	FANCM	6	5	1
14:45650900:G	A	G	16/23	c.4378A>G	p.(Ile1460Val)	missense_variant	FANCM	6	5	1
14:45665468:G	C	G	21/23	c.5434C>G	p.(Pro1812Ala)	missense_variant	FANCM	6	4	2
11:94212048:T	C	T			c.403-6G>A	splice_region_intron_variant	MRE11	5	2	3
8:30999280:T	G	T	26/35	c.3222G>T	p.(Leu1074Phe)	missense_variant	WRN	5	3	2
1:242042301:A	G	A	13/16	c.1765G>A	p.(Glu589Lys)	missense_variant	EXO1	4	3	1
16:89836323:T	C	T	26/43	c.2426G>A	p.(Gly809Asp)	missense_variant	FANCA	4	2	2
16:89849480:T	C	T	16/43	c.1501G>A	p.(Gly501Ser)	missense_variant	FANCA	4	2	2
2:29416366:C	G	C	29/29	c.4587C>G	p.(Asp1529Glu)	missense_variant	ALK	4	3	1
3:10088266:T	G	T	15/43	c.1137G>T	c.1137G>T(p.(Val379=))	splice_region, synonymous_variant	FANCD2	4	1	3
3:10140671:A	G	A	43/43	c.*37G>A		3_prime_UTR_variant	FANCD2	4	2	2
3:10140696:G	A	G	43/43	c.*62A>G		3_prime_UTR_variant	FANCD2	4	2	2
5:79960955:A	G	A			c.359-7G>A	splice_region_intron_variant	MSH3	4	3	1
8:30999123:A	G	A			c.3138+7G>A	splice_region intron_variant	WRN	4	3	1
8:90958530:C	T	C			c.1915-7A>G	splice_region_intron_variant	NBN	4	2	2
8:90990479:G	C	G	5/16	c.553G>C	p.(Glu185Gln)	missense_variant	NBN	4	2	2
1:115256529:C	T	C	3/7	c.182A>G	p.(Gln61Arg)	missense_variant	NRAS	3	1	2
12:25362777:G	A	G	6/6	c.*73T>C		3_prime_UTR_variant	KRAS	3	2	1
12:25380275:G	T	G	3/6	c.183A>C	p.(Gln61His)	missense_variant	KRAS	3	2	1
13:28610183:G	A	G			c.1310-3T>C	splice_region_intron_variant	FLT3	3	1	2
15:66782048:T	C	T			c.1023-8C>T	splice_region, intron variant	MAP2K1	3	1	2
16:23646191:C	T	C	4/13	c.1676A>G	p.(Gln559Arg)	missense_variant	PALB2	3	2	1
17:33433487:T	C	T	6/10	c.494G>A	p.(Arg165Gln)	missense_variant	RAD51D	3	2	1
17:56811608:G	C	G	9/9	c.*25C>G		3_prime_UTR_variant	RAD51C	3	2	1
2:112686988:A	G	A	2/19	c.353G>A	p.(Ser118Asn)	missense_variant	MERTK	3	2	1
2:29416481:C	T	C	29/29	c.4472A>G	p.(Lys1491Arg)	missense_variant	ALK	3	2	1
3:10106532:T	C	T	23/43	c.2141C>T	p.(Pro714Leu)	missense_variant	FANCD2	3	2	1
4:55972974:A	T	A	11/30	c.1416A>T	p.(Gln472His)	missense_variant	KDR	3	2	1
5:80168937:A	G	A	23/24	c.3133G>A	p.(Ala1045Thr)	missense_variant, splice_region_variant	MSH3	3	2	1
8:95479680:C	G	C	2/15	c.88C>G	p.(Leu30Val)	missense_variant	RAD54B	3	2	1
9:35074917:C	T	C			c.1636+7A>G	splice_region, intron variant	FANCG	3	3	0
9:98239147:G	A	G			c.1504-8T>C	splice_region, intron_variant	PTCH1	3	1	2

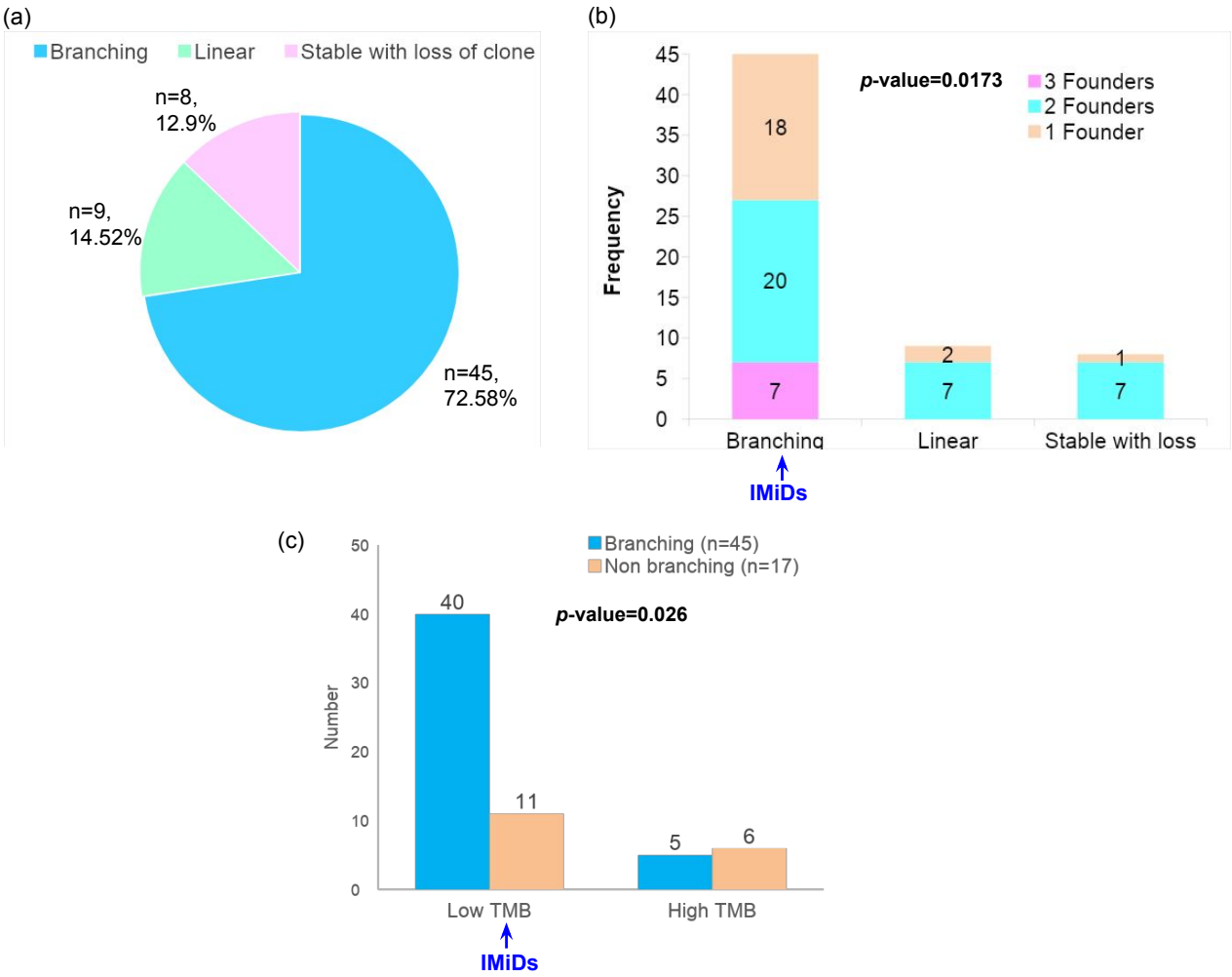


Figure 3.4: Frequencies of types of clonal evolution patterns, TMB and founder clones. (a) Distribution of types of clonal evolution patterns including branching and non branching (Linear, Stable with loss of clone) observed in MM patients, (b) Number of founder clones observed in patients with branching and non branching clonal evolution, and (c) Comparison of number of MM patients with either low or high TMB and who developed branching versus non-branching patterns of clonal evolution. Patients with branching evolution may benefit from IMiDs.

3.3.6 Heterogeneity in clonal evolution

Three types of clonal evolutionary patterns with 1 to 3 founder clones were observed in this study (Figure 3.3). The branching pattern of clonal evolution was observed in maximum number of patients (45; 72.58%) followed by Linear in 9 cases (14.51%) and Stable with loss of clone in 8 patients (12.90%) (Figure 3.3a). Distribution of founder clones in different subsets of patients with branching (n=45) and non-branching (n=17) evolution is shown in Figure 3.3b. One, two and three founder clones were detected in 18, 20 and 7 patients respectively out of 45 patients with branching patterns of clonal evolution. Patients with branching pattern of evolution had significantly higher number

of founder clones (p=0.0173, Figure 3.3b) than those with non branching patterns. A significant number of patients with low TMB at TP1 developed branching clonal evolution (n=40 out of 51) whereas those with high TMB had both branching (n=5 out of 11) and non-branching evolutionary patterns (n= 6 out of 11) (p=0.026) (Figure 3.3c).

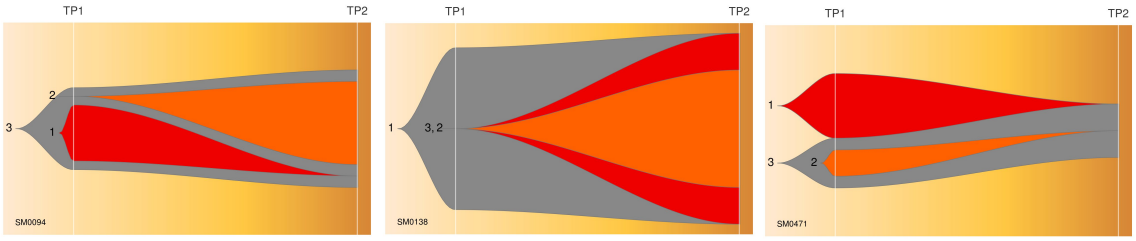


Figure 3.5: Three patterns of clonal evolution. A representative scheme of fish plots corresponding to three patterns of clonal evolution (a) Branching, (b) Linear, and (c) Stable with loss of clone

Each case of MM was analyzed in depth by QuantumClone and their individual fish plots, clonal density and evolution plots were generated (Figures B.1, B.2, B.3, B.4, B.5, B.6, B.7, B.8, B.9, B.10, B.11, B.12, B.13). A median of 3 clones (range 2 to 9) was observed among 45 patients with branching clonal evolution. The number of clones were relatively lower among patients with non-branching evolution patterns- Linear (2 to 4) and Stable with loss of clone (2 to 3). Figure 3.5a-c shows a representative fish plot of each of the three types of clonal patterns of evolution (Branching, Linear and Stable with loss) observed in this study. The somatic mutational diversity in founder clones and their cellular prevalence was compared at two time points for each patient. A schematic representation of genes found to be mutated in founder clones including actionable/non-actionable genes and the significantly associated biological pathways predicted to be affected by such mutated genes in patients are shown in Figures 3.6, 3.7 and 3.8 respectively.

The heatmaps in Figures 3.6 and 3.7 also depict falling/ rising frequencies of actionable and non-actionable targets (including DRV/ONC/TSG/others) respectively. The topmost ten genes mutated in founder clones were BAGE2 (37.28%) > *PABPC1* (30.5%) > *MUC17/ NBPF1* (23.72%) > *DNAH14/ FLG* (22.03%) > *FAT1/ RHPN2/ TPTE* (20.33%). The topmost frequently mutated actionable targets were *KRAS* (18.64%)> *BRAF/ FANCM* (13.55%) > *FANCD2/ WRN* (11.86%) > *FANCA/ MLH1* (10.16%) > *NRAS/ ATM* (8.47%) > *TET2/ BRCA1* (6.77%) > *FGFR3/ TP53* (5.08%), and others.

The cellular prevalence of topmost mutated tumor suppressor gene *KMT2C* showed an increase with progression in 6 out of 11 patients followed by FAT1 (6 out of 12), *FANCA* (3 out of 6), *BRCA1* (3 out of 4), *TET2* (2 out of 4) and *NRAS* (4 out of 5) (Figures 3.6

& 3.7). On the contrary, cellular prevalence of mutated driver *PABPC1* decreased with progression in 13 out of 18 patients, *KRAS* (8 out of 11), *BRAF* (6 out of 8), *ATM* (4 out of 5) and others (Figures 3.6 & 3.7).

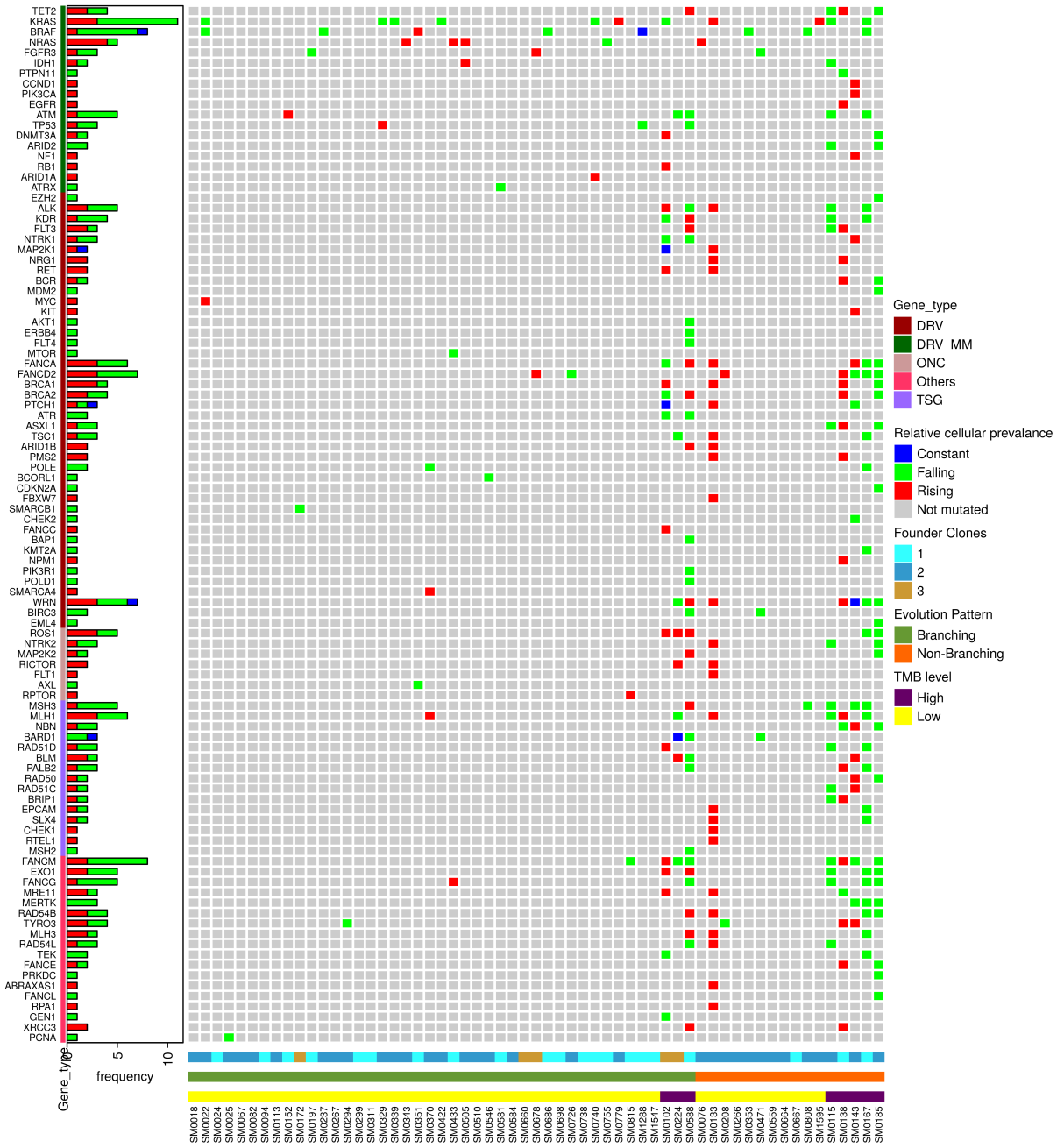


Figure 3.6: Comparison of potential actionable mutated genes in different samples grouped as with branching or non branching clonal evolution patterns and low or high TMB levels. Heatmap depicting distribution of actionable targets including drivers, oncogenes and tumor suppressors with rising or falling frequency trends across MM patients classified on the basis of branching/ non branching clonal evolutionary patterns, TMB levels and number of founder clones.

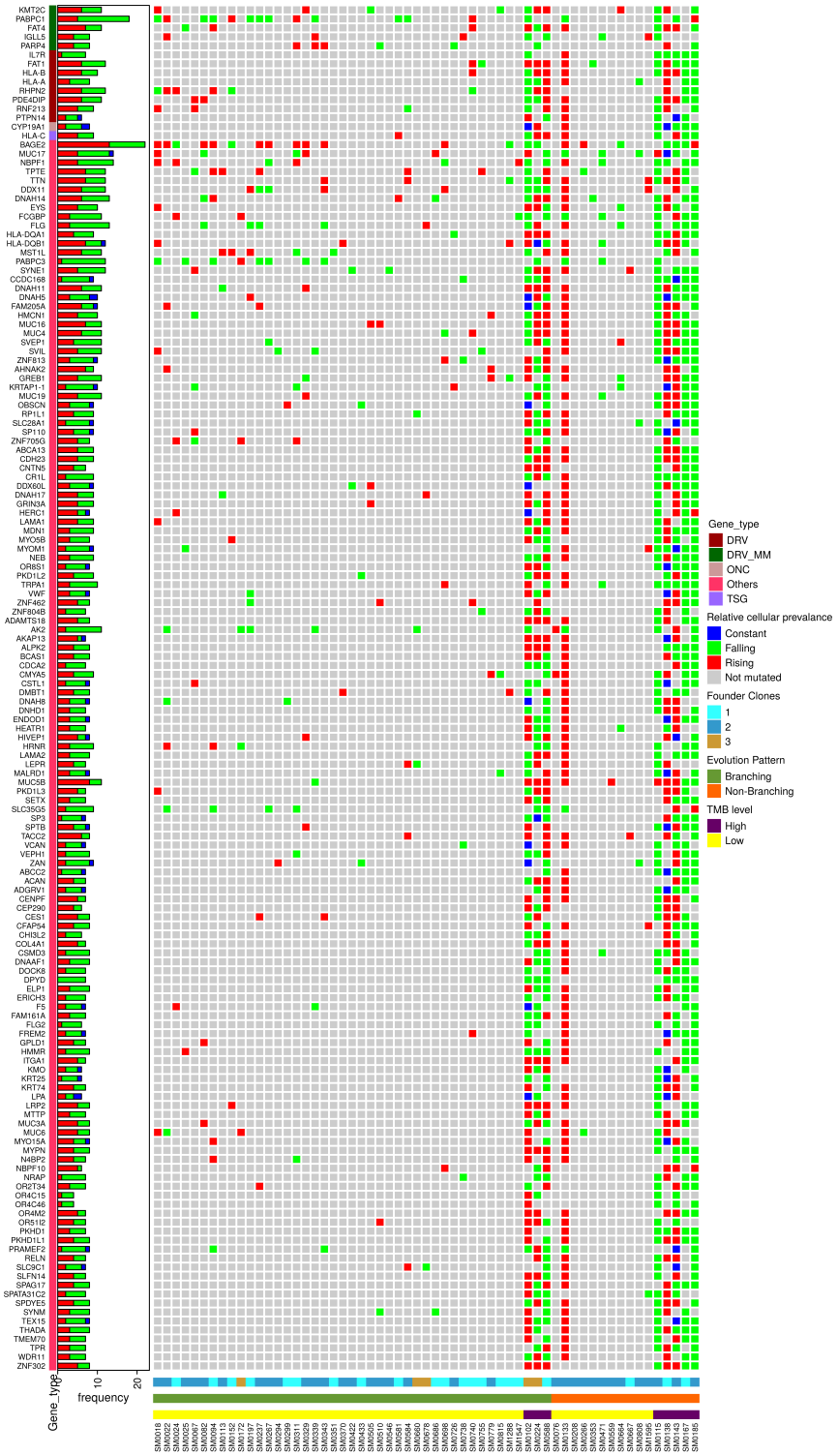


Figure 3.7: Heatmap depicting distribution of non actionable target genes drivers, oncogenes and tumor suppressors with rising or falling trends across MM patients classified on the basis of branching/ non branching clonal evolution patterns, TMB levels and number of founder clones.

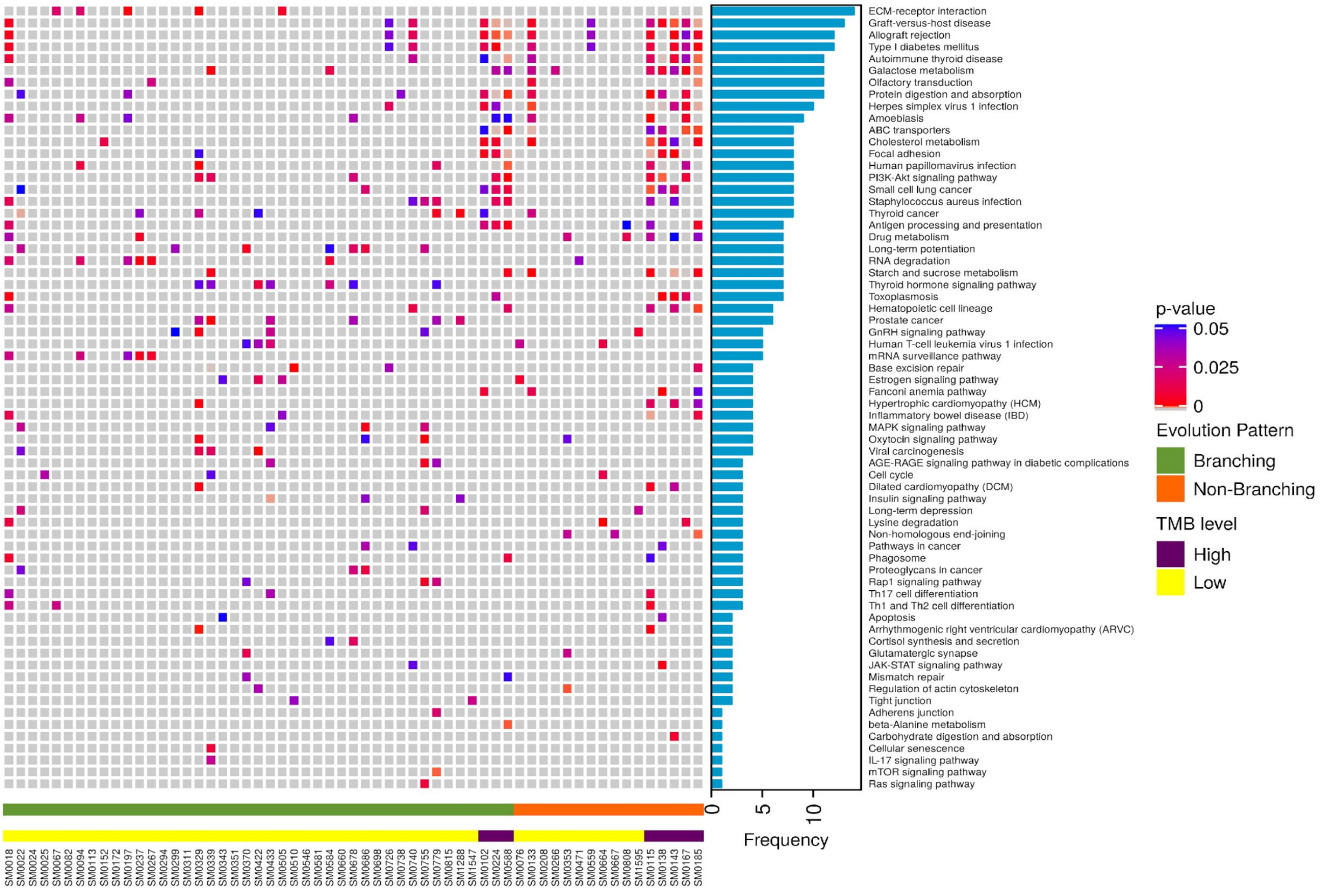


Figure 3.8: Predicted pathways affected by somatic mutations across samples. Heatmap depicting significantly affected biological pathways predicted to be altered by Enrichr across MM patients classified on the basis of branching/ non branching clonal evolutionary patterns and TMB levels

3.3.7 Prediction of biological pathways affected by somatic mutations

A comprehensive gene enrichment analysis by Enrichr identified a network of biological pathways found to be significantly associated with somatic mutations on progression of MM (Figure 3.8). These included, notably, ECM-receptor interaction, Galactose metabolism, Protein digestion and absorption, Cholesterol metabolism, Antigen processing and presentation, Drug metabolism, RNA degradation, Starch and sucrose metabolism, Hematopoietic cell lineage, Base excision repair, MAPK signaling pathway, viral carcinogenesis, cell cycle, apoptosis, Th17 cell differentiation, Th1 and Th2 cell differentiation, beta-Alanine metabolism, cellular senescence and others.

Pathways that were affected by 2434 mutated genes found exclusively at diagnosis and those affected by new mutations in genes at TP2 are shown in Figure 3.9. Additional

pathways (n=13) found to be affected exclusively on progression included NK cell mediated cytotoxicity, chemical carcinogenesis, PI3K-Akt signaling, phototransduction, PPAR signaling, GnRH signaling and others. Likewise, 18 pathways were exclusively affected by mutations at TP1.

3.3.8 Clonal divergence in individual cases

Figure 3.5a-c shows a representative fish plot of each of the three types of clonal patterns of evolution (Branching, Linear and Stable with loss) observed in this study. A case-wise description of subclones and their patterns of evolution are summarized in the Figures B.1, B.2, B.3, B.4, B.5, B.6, B.7, B.8, B.9, B.10, B.11, B.12, B.13) and Supplementary Notes (Appendix B).

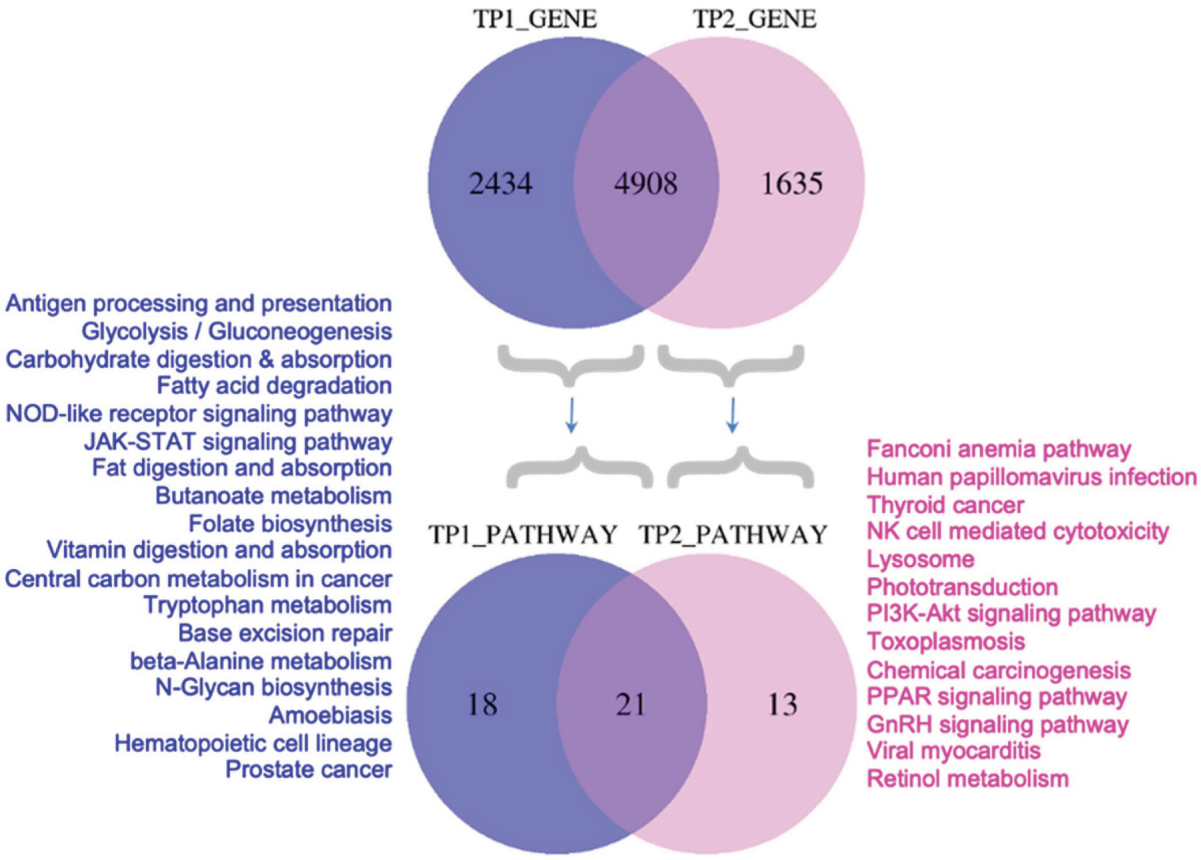


Figure 3.9: Comparison of mutated genes and associated pathways at diagnosis and at progression. Venn diagram representing number of mutated genes and the predicted biological pathways affected by mutations exclusively at diagnosis (TP1) or progression (TP2).

3.4 Discussion

Progression of MM is linked with a spatiotemporal shift in subclonal structure. The prime objective of this study was to explore subclonal evolution associated with progression of MM and identify potential actionable targets for each patient. In order to achieve this, we adopted a novel Ensemble algorithm approach for identification of mutations. As per our findings and as suggested by others [144, 145], there can be significant differences in the SNV outputs processed by different variant callers based on the properties of the caller used, their strengths and weaknesses. Since no somatic caller has the ultimate ability to perform, an ensemble approach that combines multiple callers has been reported to offer the best balance of both sensitivity and specificity [145, 146, 147]. Hence, we decided to call mutations through four common variant callers (Dragen, Strelka2, SomaticSniper and SpeedSeq) and generate a common consensus rather than depending on any single one. This innovative approach ensured that the clonal landscape of MM captured in our study was closest possible estimation to reality.

An important observation of this study is that we have been able to identify recurrent subclonal shifts in actionable/ druggable targets of clinical importance such as BRAF, KRAS, ATM, TET2 and TP53 at diagnosis in multiple patients (in atleast 3 patients or more) (Figure 3.3a). A similar gain in subclonal NRAS mutations was observed at the time of progression (Figure 3.3c). The reduction in frequencies of driver genes with progression can be explained by their selective loss in response to therapy that may coincide with fulfillment of their initial functional role(s) needed in triggering myelomagenesis. On the other hand, an increase in another set of driver genes indicates an effect of evolutionary pressure that allows selection of topmost fit clones. These sweeping subclones may either be novel or may result from expansion of pre-existing mutations known to be present at low or undetectable frequencies at the time of diagnosis or earlier. The inability to detect low copy mutations is largely due to technical limitations of sequencing of bulk tumor tissue and recent advanced technologies of single cell sequencing may be able to resolve effect of evolving somatic mutations more lucidly.

Screening of actionable genetic mutations in these genes allows to match patients with future treatments that would be most beneficial, which is in coherence with the overall goal of the ongoing Multiple Myeloma Research Foundation (MMRF) MyDRUG (Myeloma-Developing Regimens Using Genomics) clinical trial (NCT03732703) [148]. The MyDRUG aims at enrolling patients with mutations in BRAF, NRAS, KRAS, FGFR3, CDKN2C, IDH2 or t(11;14) and assign to appropriate targeted agent against that mutation. Patients with BRAF V600E or any NRAS or KRAS actionable mutations

found in subclonal populations could thus benefit the most if treated early with BRAF inhibitor e.g. Vemurafenib or MEK inhibitor Cobimetinib respectively. Heat maps in Figure 3.6 show genomic signatures of actionable genes for each patient enrolled in this study that could be targeted specifically to select the right drug for the right patient based on the specificity of the mutation.

TMB is an emerging prognostic biomarker of response to immunotherapy, approximation of neoantigen load and overall survival especially in solid tumors [149, 150]. A high TMB is considered a biomarker of higher neoantigen load, increased response rates to immunotherapy and better outcomes. High somatic mutation and neoantigen loads have been found to correlate with reduced PFS in MM [151]. Patients were classified in this study into those with low TMB between $\leq$1 to 10 or high TMB ($\geq$10 or hypermutators). This study has shown a modest loss of TMB from diagnosis to progression but only in a subset of patients with hypermutator status (i.e. TMB$\geq$10) (Figure 3.2). There could be a selective loss of less fit drug sensitive clones yet with persistence of drug resistant clones in such patients and hence combination of IMiDs with novel therapeutics could be used to treat such patients.

This study has shown a predominance of branching pattern of clonal evolution in MM in concurrence to other studies [63, 68, 69, 70, 71, 72, 73, 74] (Figure 3.4a). An increase in DNA damage and a branching pattern of evolution are considered hallmarks of effectiveness of therapy and attainment of deep response [68]. Although the branching type of evolution reflects on the better response rates to therapy while tumor strives to mutate and acquire fitter clones to survive, it is also a prominent underlying mechanism of relapse. While mutations in founder clones are primarily involved in initiation of myelomagenesis, those in subclones may contribute significantly to relapse. The study has further shown that branching evolution is more predominant among patients with 2 or more founder clones (Figure 3.4b) and those with low tumor mutation burden (TMB<10) (Figure 3.4c). Since, this happens under the positive selection pressure of therapy and the microenvironment, such patients could perhaps benefit more from immunomodulatory drugs (IMiDs) such as thalidomide/ lenalidomide and analogues [67].

Studies have shown that ongoing DNA damage intensifies from MGUS to MM and provides a mechanism by which chromosomal aberrations and heterogeneity are acquired by malignant plasma cells [152]. Figures 3.8 and 3.9 show the functional pathways that were affected by genetic mutations on progression. These include pathways in cancer, metabolism of galactose, cholesterol, drugs, cellular senescence, cell cycle, apoptosis, viral carcinogenesis, RNA degradation, base excision repair and several other crucial signalling pathways involved in pathogenesis of MM or immune surveillance. Deregulated DNA damage repair related pathways as also seen in our study have been asso-

ciated with poor prognosis [153] since the tumor cells can withstand DNA damaging drugs and repopulate with therapy resistant cells on treatment. It has been suggested that a 'synthetic lethality' approach [154] may be more beneficial where co-treatment of patients with current drugs and those targeting DNA repair pathways [155] (e.g, Bortezomib with PARP1 inhibitor [156] or Spironolactone [157] or a novel compound DCZ3301 [158]) may reverse drug resistance in such patients [159, 160].

Studies like this have shown genomic plasticity of mutational landscapes and how relative preponderance of mutated drivers changes with disease progression. Figures B.1, B.2, B.3, B.4, B.5, B.6, B.7, B.8, B.9, B.10, B.11, B.12, B.13 show individual evolution patterns as FISH plots followed by summarized individual case reports on 62 newly diagnosed MM patients enrolled in this study. It provides a detailed genomic architecture and cellular prevalence of each and every subclone identified for every patient at diagnosis and at progression. Table 3.4 summarizes the number of patients who had an actionable/ druggable mutation and who could qualify for targeted treatments with target specific drugs. Comprehensive analysis of mutational subclonal landscapes of patients as observed in this study are pre-requisites to infer the genomic mutations that can be treated in future in similar lines as in MyDRUG trial. An integration of such early genomic biomarkers with clinical biomarkers could help in risk estimation and identification of patients who could benefit more from a rationalized therapeutic approach at early stages. It is indeed not just the individual mutations but an extended treatment landscape that needs to be monitored preferably at multiple time points to tailor therapy. An early assessment of TMB along with mutations in drivers and actionable target genes during decision making, may therefore, allow most appropriate therapeutic personification in clinics.

3.4.1 Conclusion

This study explored the subclonal evolution associated with the progression of MM and identified the potential actionable targets for each patient. A marked intraclonal heterogeneity was observed in all the patients and the disease progression was characterized by recurrent subclonal shifts in the actionable targets such as KRAS, ATM, TET2 and TP53 while gain in subclonal mutations in NRAS. Based on the specificity of the actionable driver mutations revealed by the temporal analysis of the variants at the two time points, appropriate drug for the individual patients could be selected thereby leading to personalized treatment. Further, the genomic mutations in addition with clinical biomarkers could help in identification of high risk patients who are still in the initial stages of the disease. Timely medical intervention could be provided to such high risk patients to slow down the progression of disease and improve their overall survival.

Branching pattern of evolution was observed among 72.58% patients and was found to be more predominant in patients with low TMB (64.51%) had(<10) and 2 or more founder clones (61.29%).

This study also revealed loss in the TMB from diagnosis to progression in hypermutator patients who may benefit from IMiDs. However, it needs to be validated on a larger cohort of patients. Thus, a systematic analysis of evolving mutational landscapes at multiple time points in addition with TMB and SBS signatures could help in better stratification of high risk MGUS/SMM/MM patients prior to subclonal expansion and therefore open the opportunities of early and personalized cure for the disease. MGUS and SMM are both precursor stages to MM. While MGUS and SMM are both benign conditions, displaying no clinical symptoms, there is a higher risk of progression to MM in SMM patients than MGUS patients. Therefore, genomic landscape of MGUS and SMM patients should be studied in conjunction with MM to identify the distinctive features that ultimately leads to MM. It was difficult to collect exome data of MGUS and SMM as these are non-malignant stages but we were able to get access to MGUS data of 61 patients. The comparative study of MGUS and MM patients has been presented in chapter 4 in detail.

Chapter 4

Mutational landscape of MM and its precursor MGUS

4.1 Introduction

MGUS being a precursor of MM shows a genetic profile which is similar to MM, however, overall the mutations are present at a lower level as compared to MM. This indicates that there are additional mutations taking place in MGUS genome over time which finally leads to MM. Therefore, in this study, we have studied the exome data of MGUS and MM patients to reveal the entire spectrum of mutations altered in MGUS and MM and how it evolves over time. We explored the change in the mutational landscape as the disease progressed from the MGUS to MM. We found that the difference in the frequency of the single base substitution is significantly different in MGUS and MM. We have also analyzed the frequency of the different types of variants across MGUS and MM and found that few have changed significantly as the disease progressed from MGUS to MM. Further, we categorized MM patients into low TMB and high TMB (hypermutators) based on their overall survival data. We explored the impact of TMB on the frequency of single base substitutions and the different variant types across the low and high TMB groups of MM patients. The association of TMB with overall survival is still unknown in newly diagnosed multiple myeloma (NDMM) patients; therefore, we have correlated TMB with survival data and found that high TMB is linked with poor overall survival in NDMM patients.

4.2 Materials and Methods

4.2.1 Datasets used in the study

The present study is based on the data of 1018 NDMM patients and 61 MGUS patients. Variant files generated from the exome data of 936 NDMM patients out of the total 1018 patients were obtained from the GDC portal via dbGaP authorized access (phs000748; phs000348). This data is a part of the MMRF CoMMpass study. Exome data of the remaining 82 NDMM patients were obtained from AIIMS, Delhi. In addition, exome data of 33 MGUS patients out of 61 patients was obtained from EGA (EGAD00001001901), and exome data of the remaining 28 patients was obtained from AIIMS, Delhi. Four variant callers, namely, MuSE [19], Mutect2 [20], VarScan2 [18],

and Somatic-Sniper [21], was used for finding variants in patients from the MMRF CoMMpass study. Therefore, there were four vcf files corresponding to each variant caller for each patient. The workflow of the complete analysis is shown in Figure 4.1.

4.2.2 Analysis of exome data and the variants identified using the exome data

Exome data obtained from AIIMS and EGA was processed with a standard exome sequencing pipeline, and single nucleotide variants (SNVs) were extracted using MuSE, Mutect2, VarScan2, and Somatic-Sniper variant callers. SNVs were annotated using ANNOVAR [26] to gather the genomic information of the mutations, such as their variant type and the deleteriousness of the mutation, etc. FATHMM-XF [31] was used to remove the benign variants. The rest of the filtered variants were categorized into nonsynonymous (NS) variants, synonymous (SYN) variants, and other (OTH) variants. Exonic, nonsynonymous single nucleotide variants (snvs), ncRNA_exonic, stop gain, stop loss, start loss, splicing, frameshift insertion, and frameshift deletion were grouped in nonsynonymous variants. UTR3, synonymous single nucleotide variants (snvs), and UTR5 were grouped in synonymous variants. Non-frameshift insertion, non-frameshift deletion, non-frameshift substitution, intronic, intergenic ncRNA_intronic, upstream, downstream, unknown, and ncRNA_splicing were grouped in other variants.

4.2.3 Assessment of single base substitution, mutational signatures, and TMB

Variants identified by three or more callers were further processed to extract information on single base substitution and identify the mutational signatures present in the data. SigProfilerExtractor [161] was used to discover the single base substitutions and the mutational signatures in the MGUS and MM data. The etiology of the deduced signatures were found via the COSMIC v3.2 mutational signature database [162]. A total of six single base substitutions C>A, C>G, C>T, T>A, T>C, and T>G were identified. Tumor mutational burden (TMB) was calculated using the three different categories of variants- nonsynonymous (NS) variants, synonymous (SYN) variants, and other (OTH) variants. TMB was determined as described in [163]. TMB_NS, TMB_SYN, and TMB_OTH were estimated using nonsynonymous (NS) variants, synonymous (SYN) variants, and other (OTH) variants, respectively. Survival data were available for 832 (753+79) patients out of a total of 1018 NDMM patients, which were utilized to obtain the threshold values for TMB_NS, TMB_SYN, and TMB_OTH using the K-adaptive partitioning (KAP) algorithm [164] and Cutoff Finder [165].

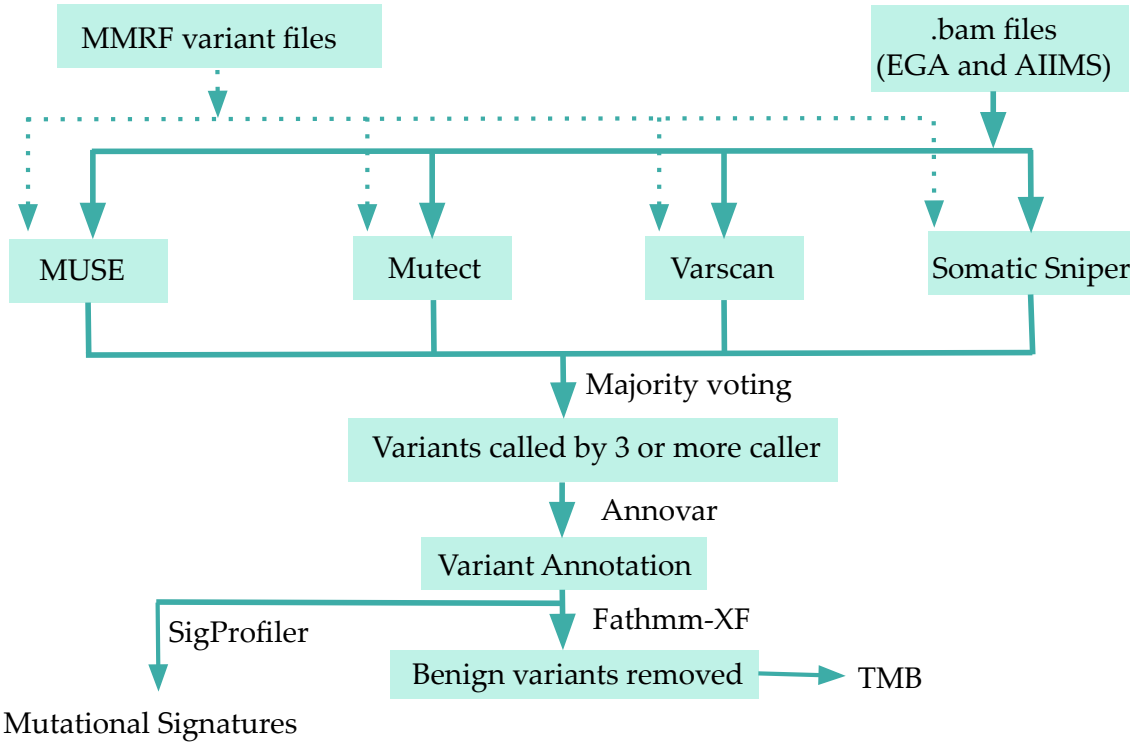


Figure 4.1: Workflow of the study and data analysis. Four different variant callers were used to identify variants in the MM and MGUS patients. Variants were finalized using the majority voting scheme. Variants were then annotated with Annovar for deducing TMB. Mutational signatures were inferred using Sigprofiler tool.

4.2.4 Statistical analysis

Wilcoxon rank-sum test was used to determine if the change in the frequencies of the single base substitutions and the different types of variants is statistically significant between the MGUS and MM. Unpaired Wilcoxon rank-sum was applied because the data did not follow the normality distribution and was unpaired.

4.3 Results

4.3.1 Frequency of single base substitutions (SBS) increases significantly from MGUS to MM

There was an increase in the median and mean frequency of the single base substitutions from MGUS to MM. The change in the frequency was statistically significant with p-values less than 0.05 for all the substitutions according to the Wilcoxon rank-sum test (Figure 4.2). C>T substitution was observed with the highest frequency in MGUS and MM, increasing the median value from 30 to 59. T>C substitution was next, with an increase in the median value from 20 (MGUS) to 35 (MM). T>A was observed with the lowest frequency in MGUS and MM, increasing the median value from 7 to 17.

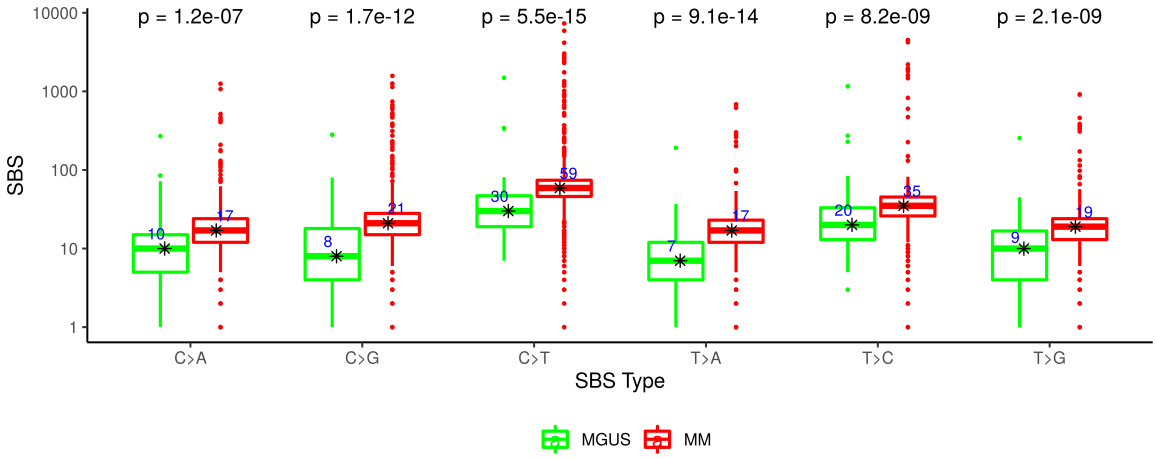


Figure 4.2: Boxplot shows the difference in the frequency of the single base substitutions between MGUS and MM patients. Wilcoxon rank-sum test was applied to determine if the change is statistically significant or not. For all the substitutions, there is significant variation in the frequency with p-values less than 0.05 between the two groups.

4.3.2 Calculation of threshold values for the SBS and comparison between the high and low-frequency MM groups

Due to the availability of survival data for 832 MM patients, threshold values for the substitutions were inferred. K-adaptive partitioning (KAP) algorithm and Cutoff Finder were used to deduce the thresholds. Table 4.1 shows the cut-off values estimated for the different types of substitutions for PFS and OS via KAP while Figures 4.3, 4.4, 4.5 and 4.6 show the cut-offs deduced via Cutoff Finder for PFS and OS respectively. The higher of the two cut-offs obtained via KAP were selected for C>T, T>C, C>G,

Table 4.1: The table shows the cut-offs obtained for the six different types of substitutions via KAP. Two cut-offs were obtained for each SBS, one using PFS and the other using OS. The higher of the two cut-offs and the patients were then organized into two groups, one with SBS values less than the selected cut-offs and the other one with SBS values greater than the selected cut-offs. KM analysis showed that there was a significant difference in the survival patterns of the two groups of patients for the substitutions, C>T, C>G, C>A, and T>A. However, cut-offs obtained for T>C and T>G substitutions did not yield a significant difference in the survival curves. Therefore, cutoffs were manually deduced for the two substitutions where the KM curve has the maximum separability. Text in bold shows the selected cutoffs.

SBS	**Min**	**Median**	**Max**	**PFS cut-off**	**OS cut-off**	**Manual cut-off**	**Freq. ($\leq$, $>$)**	**PFS p-value**	**OS p-value**
C>A	0	17	1251	26	**28**	-	712, 120	0.00025	5.13E-06
C>G	0	21	1575	**37**	34	-	763, 69	0.026	2.20E-04
C>T	1	59	7315	79	**99**	-	750, 82	0.001	4.80E-06
T>A	0	17	684	5	**32**	-	784, 48	0.01	0.005
T>C	0	35	4498	12	11	**80**	816, 16	0.19	0.01
T>G	0	19	915	6	6	**41**	804,28	0.018	0.007

C>A, T>G, and T>A substitutions and were 99, 12, 37, 28, 6, and 32, respectively. The patients were then organized into two groups, one with SBS values less than the selected cut-offs and the other one with SBS values greater than the chosen cut-offs. Kaplan Meier (KM) curves corresponding to the two groups revealed that there was a significant difference in the survival patterns of the two groups of patients for the substitutions, C>T, C>G, C>A, and T>A. However, cut-offs obtained for T>C and T>G substitutions yielded a significantly poor outcome for the group with values less than the selected cut-offs. Therefore, cut-offs were manually deduced for T>C and T>G substitutions where the KM curve has the maximum separability and was found to be 80 and 41, respectively. Univariate and multivariate hazard analysis was also done using the selected cut-offs via KAP, as shown in the Table 4.2. The hazard ratio for all the substitutions was greater than 1 in the univariate analysis, demonstrating that an increase in the frequency of these substitutions correlated with an enhanced risk in MM patients. Univariate analysis revealed that C>T substitution had the most significant impact (p-value <0.05) on the overall survival (OS) owing to the highest hazard ratio followed by T>C and C>A while T>G had the most significant impact (p-value <0.05) on PFS followed by C>T and C>A. However, only C>A was significant in multivariate analysis with *p*-values less than 0.05 (0.04 for PFS and 0.03 for OS).

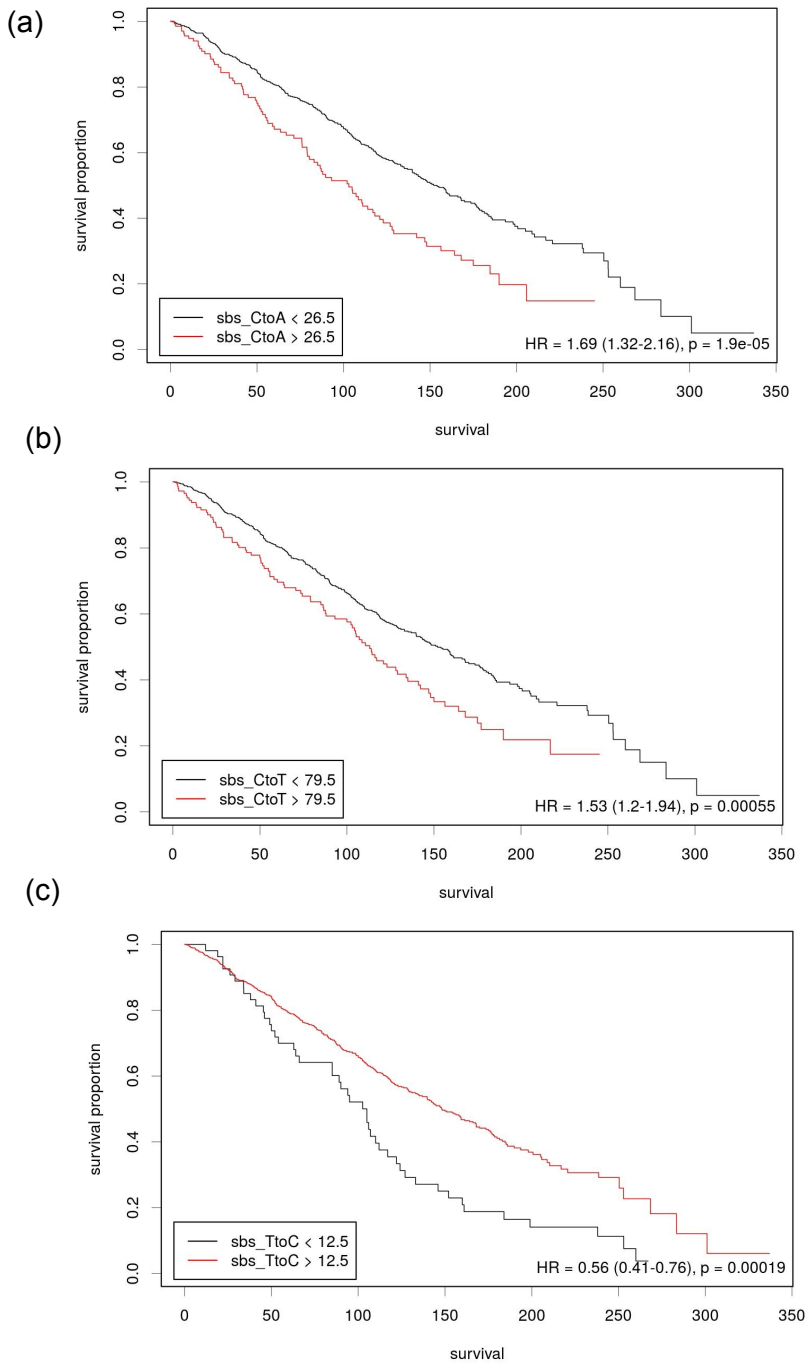


Figure 4.3: KM curves reveal differences in the PFS survival patterns of substitutions (a) C>A, (b) C>T, and (c) T>C at the thresholds obtained via Cutoff Finder. Separation in the survival curves is significant if p-values <0.05.

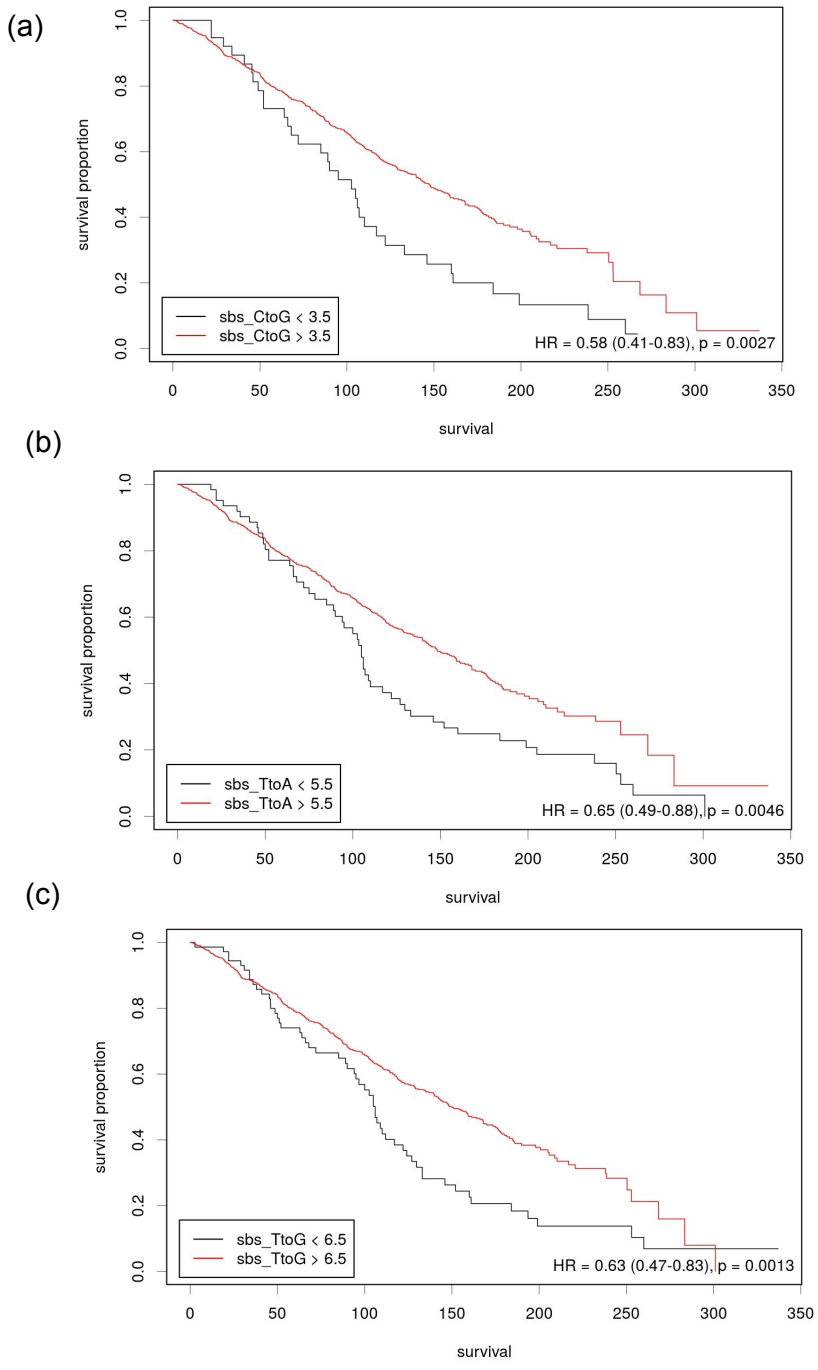


Figure 4.4: KM curves reveal differences in the PFS survival patterns of substitutions (a) C>G, (b) T>A and (c) T>G at the thresholds obtained via Cutoff Finder. Separation in the survival curves is significant if p-values <0.05.

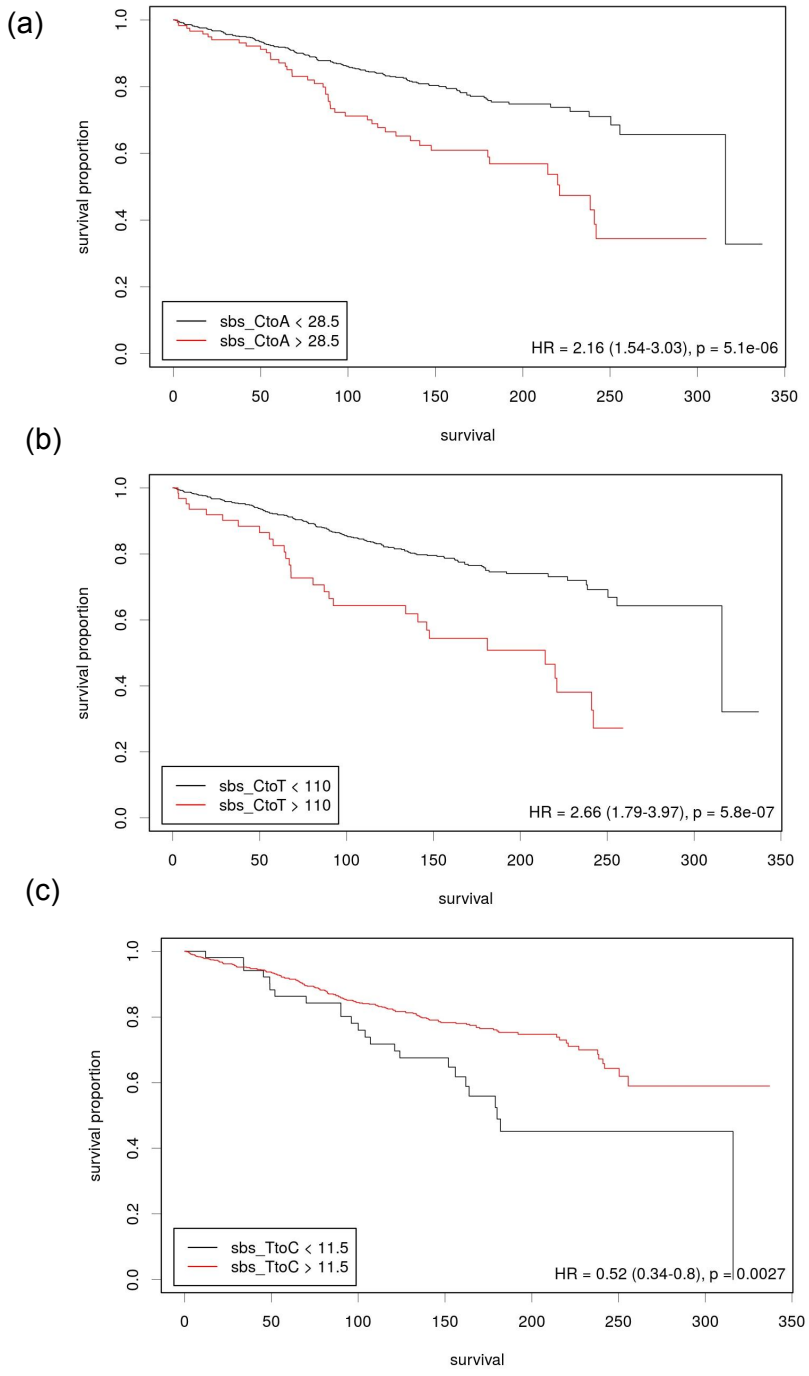


Figure 4.5: KM curves reveal differences in the OS survival patterns of substitutions (a) C>A, (b) C>T, and (c) T>C at the thresholds obtained via Cutoff Finder. Separation in the survival curves is significant if p-values <0.05.

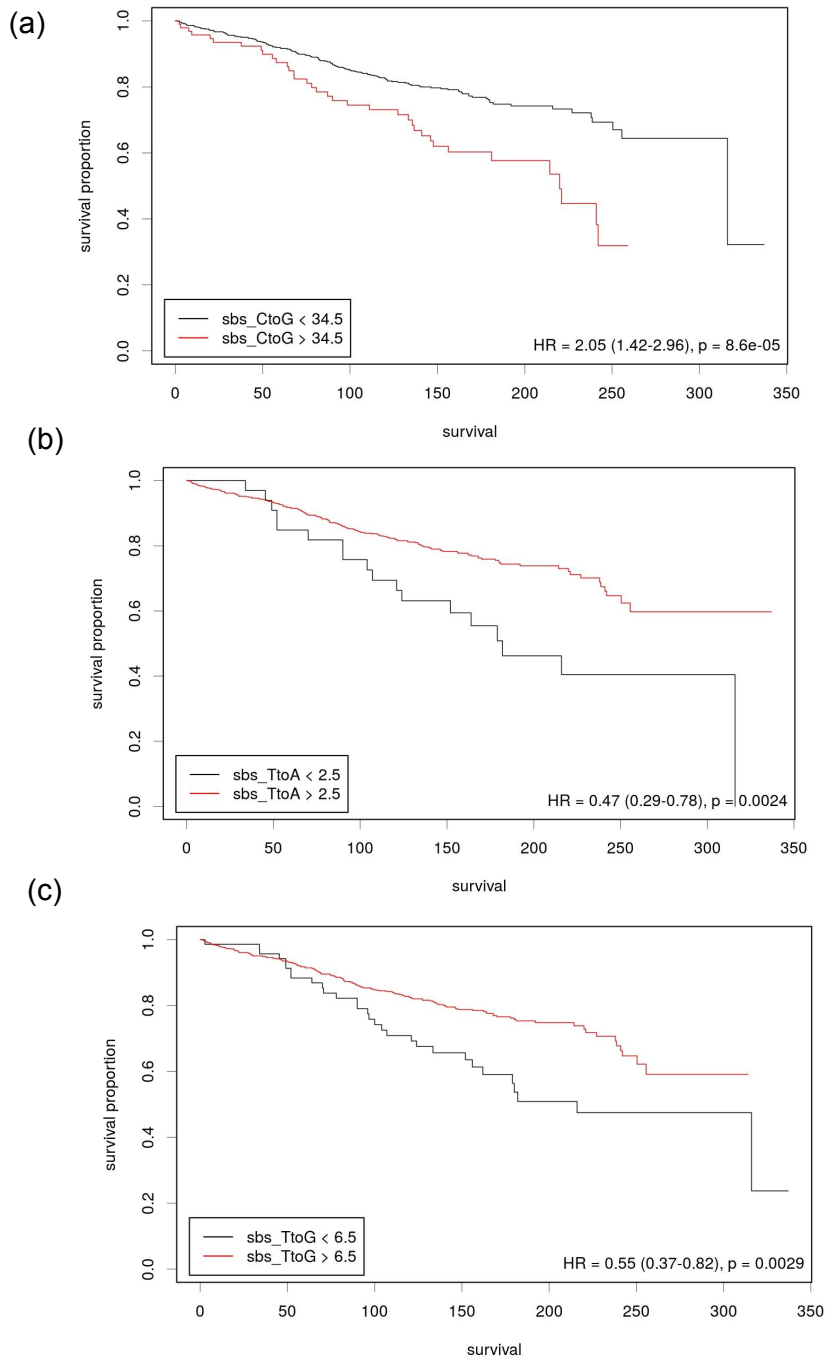


Figure 4.6: KM curves reveal differences in the OS survival patterns of substitutions (a) C>G, (b) T>A and (c) T>G at the thresholds obtained via Cutoff Finder. Separation in the survival curves is significant if p-values <0.05.

4.3.3 Comparison of mutational signature profiles between MGUS and MM

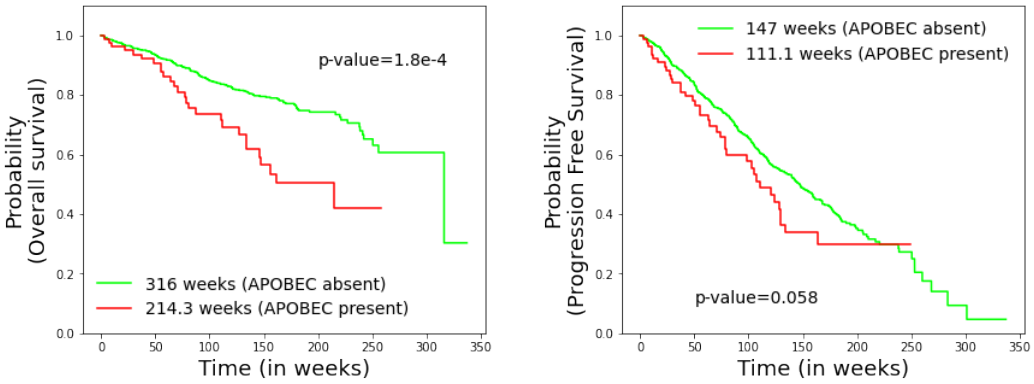
A total of 29 and 61 SBS signatures were extracted from the mutation data of MGUS and NDMM patients, respectively. Union of 29 and 61 signatures resulted in 66 unique

Table 4.2: The table shows the univariate hazard analysis and multivariate hazard analysis on the six different substitutions. T>C was removed from multivariate analysis as it was not significant for PFS in univariate analysis.

	PFS				**OS**			
	HR	**CI**	**p-value**	**C-index**	**HR**	**CI**	**p-value**	**C-index**
Univariate								
C>A	1.63	1.26-2.11	<0.005	0.54	2.16	1.54-3.03	<0.005	0.55
C>G	1.46	1.04-2.04	0.03	0.52	2.11	1.41-3.16	<0.005	0.53
C>T	1.65	1.22-2.24	<0.005	0.53	2.36	1.61-3.45	<0.005	0.55
T>A	1.61	1.11-2.32	0.01	0.51	1.93	1.20-3.11	0.01	0.52
T>C	1.47	0.83-2.61	0.19	0.5	2.27	1.19-4.32	0.01	0.51
T>G	1.73	1.09-2.75	0.02	0.51	2.14	1.21-3.77	0.01	0.51
Multivariate								
C>A	1.43	1.02-1.99	0.04	0.55	1.67	1.06-2.63	0.03	0.58
C>G	0.84	0.49-1.43	0.52		0.97	0.49-1.93	0.93	
C>T	1.38	0.86-2.22	0.18		1.71	0.91-3.22	0.1	
T>A	1.19	0.71-1.97	0.65		1.22	0.61-2.44	0.58	
T>G	1.03	0.52-2.05	0.94		0.82	0.34-1.99	0.66	

signatures. Signatures SBS37, SBS49, and SBS55 were found only in MGUS. However, their frequency is low as they were found in a single sample in MGUS (1/61=1.6%). SBS49 and SBS55 signatures are possible sequencing artifacts, and the proposed etiology of signature 37 is unknown according to the COSMIC v3.2 mutational signature database. Further, 37 signatures were discovered only in MM. However, 7 out of 37 were mutated in more than 1% MM samples. They include SBS6, SBS7d, SBS9, SBS17b, SBS19, SBS40, and SBS42. The rest of the 30 signatures were found in less than 1% MM samples and include SBS7c, SBS8, SBS10d, SBS14, SBS20, SBS21, SBS22, SBS23, SBS25, SBS26, SBS27, SBS28, SBS30, SBS32, SBS33, SBS34, SBS35, SBS36, SBS39, SBS41, SBS43, SBS46, SBS47, SBS50, SBS52, SBS53, SBS57, SBS86, SBS88, and SBS89. SBS27, SBS43, SBS46, SBS47, SBS50, SBS52, SBS53, and SBS57 are possible sequencing artifacts, as described previously. Clock-like signatures SBS1 and SBS5 were present in both MGUS and MM. Defective DNA mismatch repair signatures SBS15 and SBS44 were present in both MGUS and MM while SBS6, SBS14, SSB20, SBS21, SBS26 were present only in MM. SBS2 and SBS13 are associated with the activity of the AID/APOBEC family of cytidine deaminases and were found in both MGUS and MM. MM patients with APOBEC signatures were investigated further using survival data. APOBEC signature was present in 27 out of 177 MM patients with poor OS outcome and 52 out of 655 MM patients with superior OS outcome. Fisher's exact test revealed a statistically significant association between the APOBEC activity and poor overall survival in MM (p-value=0.0056). However, there was no significant association between APOBEC activity and progression-free survival (p-value=0.9). KM curves showed a significant difference (p-value=1.8e-4) in the over-

all survival pattern of MM patients with and without APOBEC activity (Figure 4.7). SBS84 and SBS85 are related to indirect effects of activation-induced cytidine deaminase (AID) induced somatic mutagenesis in lymphoid cells and were found in both MGUS and MM.



(a) Overall survival

(b) Progression Free Survival

Figure 4.7: KM curves reveal that APOBEC activity is associated with poor overall survival in NDMM patients. The difference in the overall survival probability between low and high TMB_NS is statistically significant with *p*-values 1.8e-4. However, there is no statistically significant difference between progression-free survival and APOBEC activity.

4.3.4 Frequency of the variants increases significantly from MGUS to MM

According to the Wilcoxon rank-sum test, there was a statistically significant increase in all the three categories of variants from MGUS to MM (Figure 4.8). The median value of nonsynonymous variants increased from 19 to 36 (p-value=5.2e-13) as the disease progressed from MGUS to MM. Median value of synonymous variants increased from 6 to 26 (p-value<2e-16) while that of other variants increased from 69 to 100 (p-value=0.007). Within the nonsynonymous category, there was a statistically significant increase in the nonsynonymous snv (p-value=2.9e-13) from 14 to 30 and stop-gain (p-value=0.016) variants from 0 to 2 as the disease progressed from MGUS to MM (Figure 4.9a). Within the synonymous category, there was a statistically significant increase in the UTR3 (p-value<2e-16) and UTR5 variants (p-value=2.7e-7) (Figure 4.9b). Within the other variant category, there was a statistically significant increase in the intronic and downstream variants (Figure 4.9c). The median value of UTR3 variants increased from 4 to 21, while that of UTR5 increased from 1 to 4.

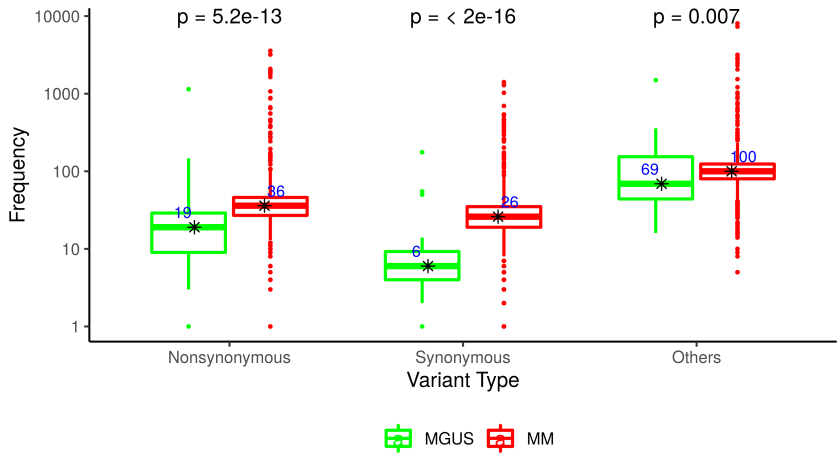


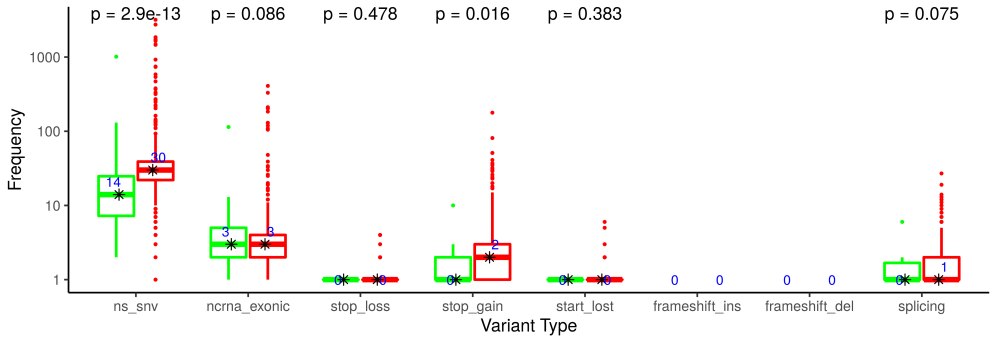
Figure 4.8: Boxplot showing the variation in the frequency of the three different categories of variants- Nonsynonymous (NS), Synonymous (SYN), and Others (OTH) between MGUS and MM. Wilcoxon rank-sum test was applied to determine if the change is statistically significant i.e. *p*-value is less than 0.05.

4.3.5 Comparison of TMB values between MGUS and MM

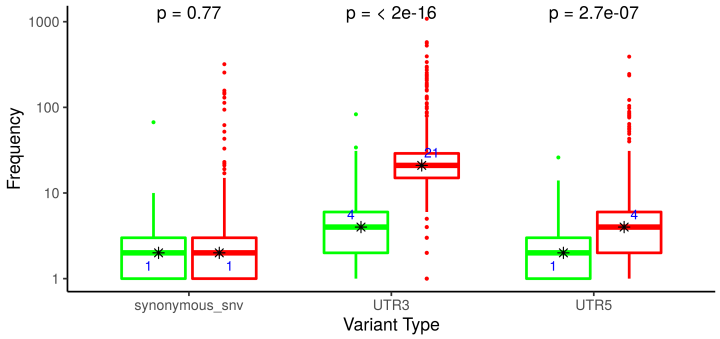
Tumor mutational burden (TMB) was calculated using the three different categories of variants- nonsynonymous (NS), synonymous (SYN), and others (OTH). A statistically significant increase was observed for TMB_NS and TMB_SYN with p-values less than 0.05 (Figure 4.10). For TMB_OTH, the difference in the KM survival curve was not significant (Figure 4.10).

4.3.6 Calculation of TMB cut-offs and comparison between high and low TMB MM groups

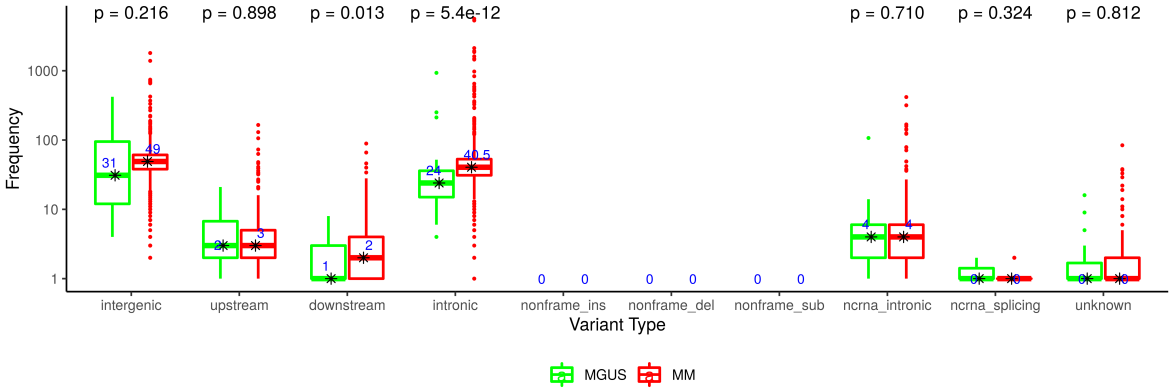
Survival data were available for 832 MM patients. Hence, threshold values of TMB were calculated using the K-adaptive partitioning (KAP) algorithm and Cutoff Finder. Both the tools inferred almost the same cut-offs (Table 4.3, Figures 4.11 and 4.12). Table 4.3 reveals the different cut-offs obtained for progression-free survival (PFS) and overall survival (OS) via KAP. For TMB_NS, 0.63 and 0.62 are the threshold values obtained via PFS and OS. Similarly, for TMB_SYN, 0.55 and 0.52 are the threshold values obtained for PFS and OS. The patients were then organized into two groups, one with TMB values less than the selected cut-offs and the other one with SBS values greater than the chosen cut-offs. There was a significant difference (p-value<0.05) on the KM survival curves of the patients below 0.63/0.62 and above 0.63/0.62. There is a significant difference (p-value<0.05) on the KM survival curves of the patients below 0.55/0.52 and above 0.55/0.52. Univariate and multivariate hazard analysis was also



(a) Change in frequency of the subcategories of nonsynonymous (NS) variants.



(b) Change in frequency of the subcategories of synonymous (SYN) variants.



(c) Change in frequency of the subcategories of other (OTH) variants.

Figure 4.9: a) Boxplot showing the variation in the frequency of the variants under the non-synonymous category. There was a statistically significant variation in the frequency of non-synonymous_snv and stop_gain variants with *p*-values less than 0.05. b) Boxplot showing the variation in the frequency of the variants under the synonymous category. There was a statistically significant variation in the frequency of UTR3 and UTR5 variants with *p*-values less than 0.05. c) Boxplot showing the variation in the frequency of the variants under the other variants category. There was a statistically significant rise in the frequency of intronic and downstream variants with *p*-values less than 0.05. Wilcoxon rank-sum test was applied to determine if the change is statistically significant or not.

done using the cut-offs via KAP, as shown in the Table 4.4. Hazard ratios for TMB_NS, TMB_SYN and TMB_OTH were greater than 1 in both the univariate and multivariate

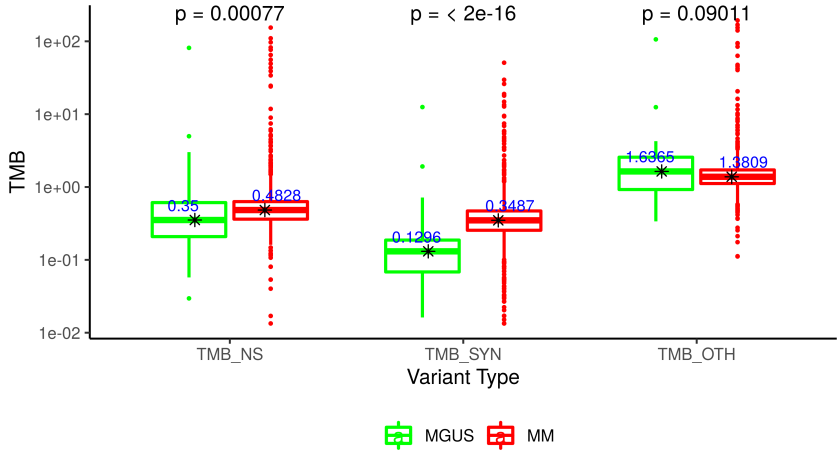


Figure 4.10: Boxplot reveals that the difference in the low TMB and high TMB groups is statistically significant with p-values less than 0.05 for TMB_NS and TMB_SYN. Wilcoxon rank-sum test was applied to determine if the change is statistically significant or not.

analysis and indicate the enhanced risk associated with an increase in the mutation burden. Multivariate analysis showed the combined effect of the TMB values on the survival patterns where TMB_NS had the highest impact, followed by TMB_OTH and TMB_SYN, respectively.

Table 4.3: The table shows the cut-offs obtained for TMB_NS, TMB_SYN aand TMB_OTH via KAP. Two cutoffs were obtained, one using PFS and the other using OS. The two cut-offs obtained for TMB_NS and TMB_SYN are close to each other. The same cut-off was obtained using PFS and OS for TMB_OTH. There was a significant difference (p-value < 0.05) on the KM survival curves of the patients below and above the selected cut-offs.

	Min Median Max	**KAP on PFS**			**KAP on OS**		
		Cut-off (<=,>)	**PFS**	**OS**	**Cut-off (<=,>)**	**PFS**	**OS**
TMB_NS	0 0.496 154.2	0.63 (612, 220)	3.19E-07	3.52E-08	0.62 (611, 221)	3.90E-07	2.09E-08
TMB_SYN	0 0.3487 50.84	0.55 (703, 129)	4.12E-05	2.05E-08	0.52 (668, 164)	5.60E-04	3.50E-08
TMB_OTH	0.1114 1.3742 193.673	1.84 (666, 166)	4.90E-06	9.16E-09	1.84 (666, 166)	4.90E-06	9.16E-09

MM patients with very high TMB_NS load and very low TMB_NS load were analyzed separately. Cut-off of 35 and 0.1 was deduced using the maximum separability on the KM survival curves. There were 822 patients with TMB_NS less than 35 and only 10 with TMB_NS greater than 35. There were 6 patients with TMB_NS less than 0.1 and 826 patients with TMB_NS greater than 0.1. A significant difference in the survival

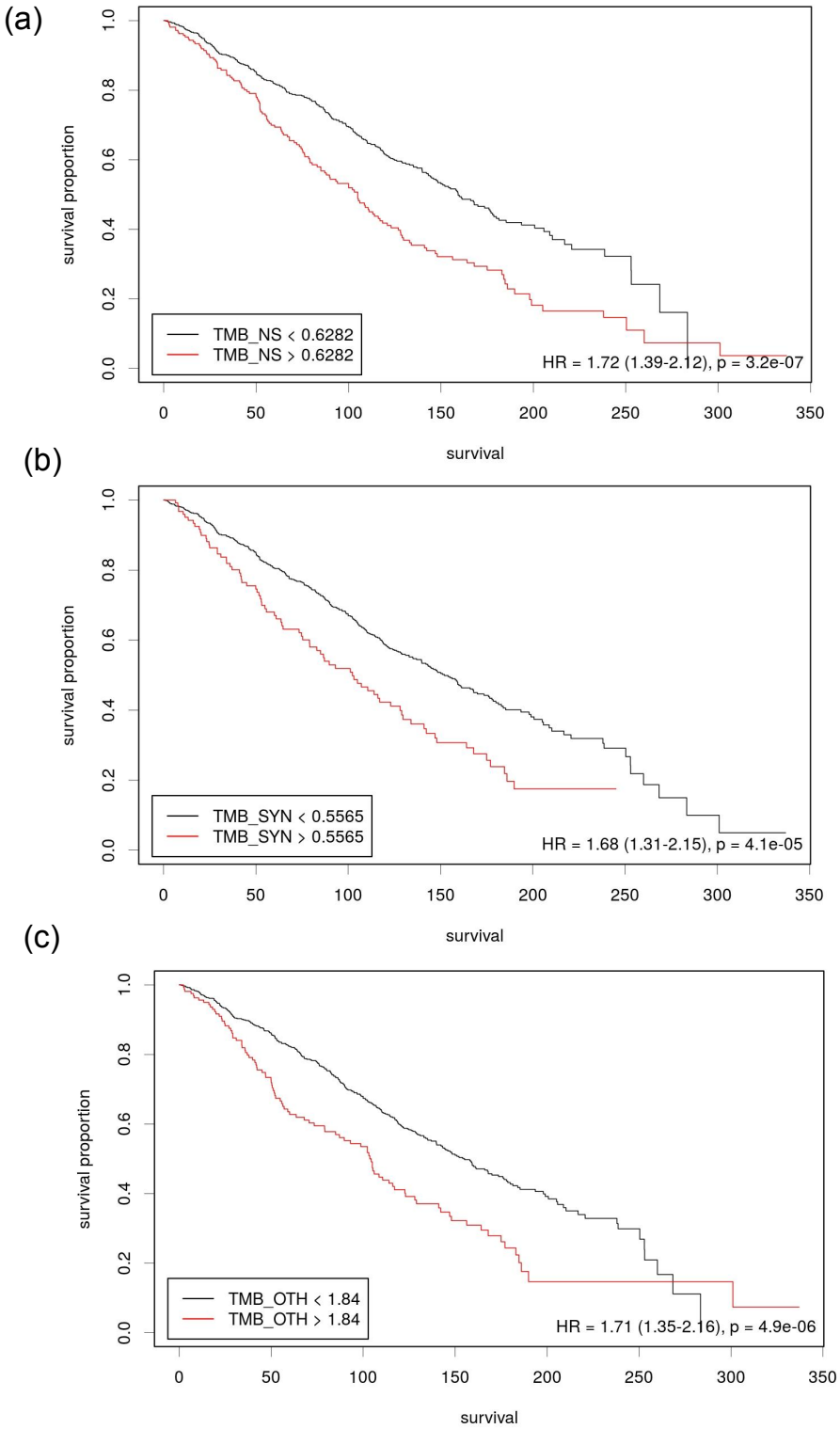


Figure 4.11: KM curves reveal differences in the PFS survival patterns of different categories of TMB (a) TMB_NS, (b) TMB_SYN, and (c) TMB_OTH at the thresholds obtained via Cutoff Finder. Separation in the survival curves is significant if p-values <0.05.

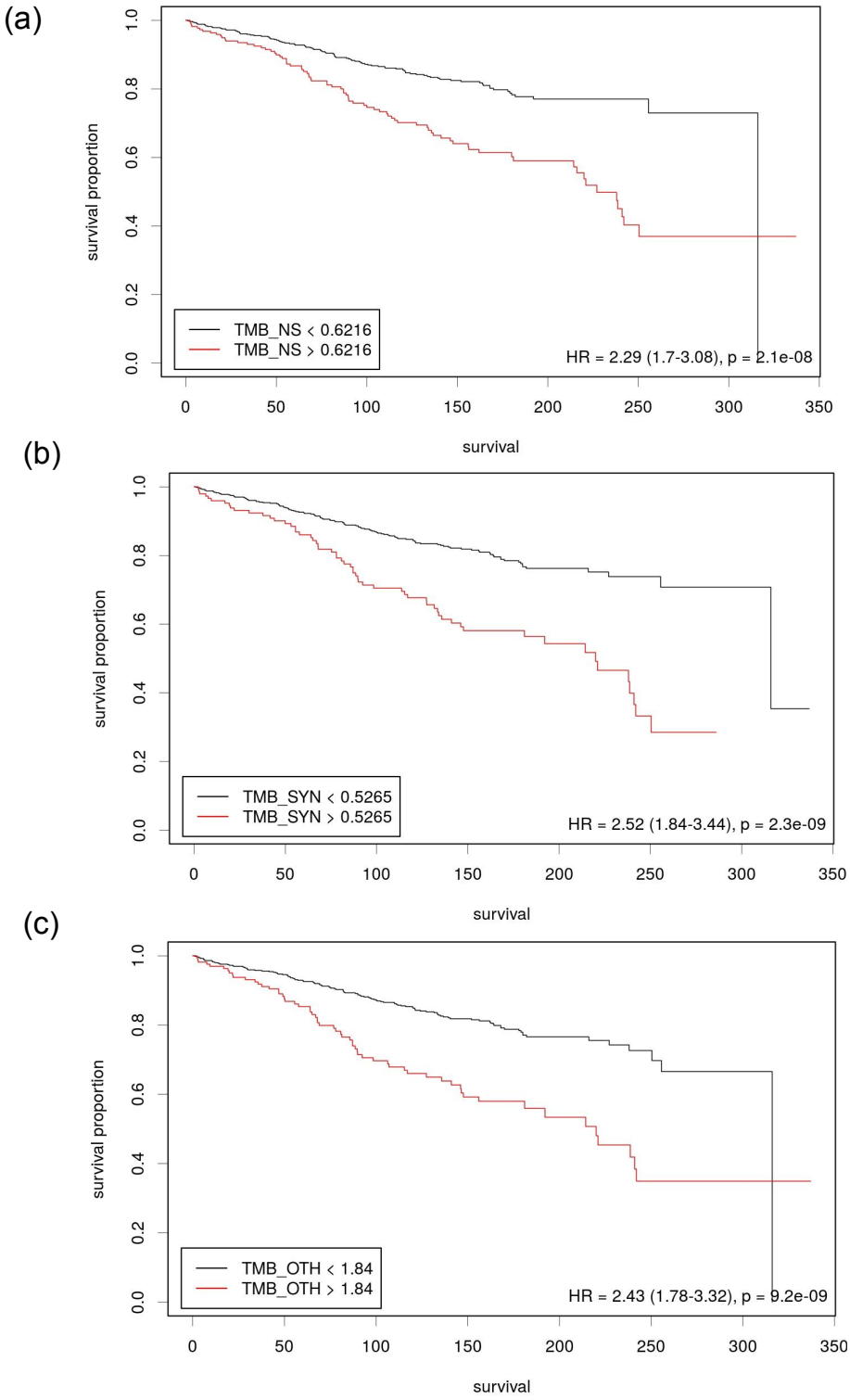
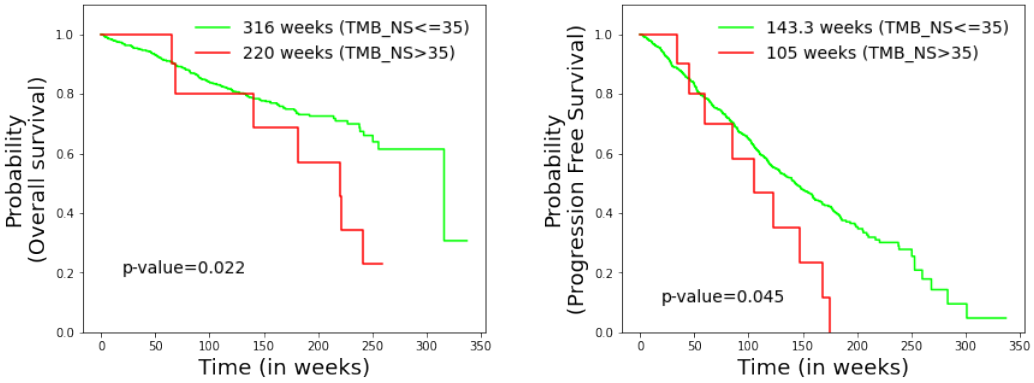


Figure 4.12: KM curves reveal differences in the OS survival patterns of different categories of TMB (a) TMB_NS, (b) TMB_SYN, and (c) TMB_OTH at the thresholds obtained via Cutoff Finder. Separation in the survival curves is significant if p-values <0.05.

Table 4.4: The table shows the univariate hazard analysis and multivariate hazard analysis obtained on TMB_NS, TMB_SYN and TMB_OTH.

	PFS				OS			
	HR	**CI**	**p-value**	**C-index**	**HR**	**CI**	**p-value**	**C-index**
Univariate								
TMB_NS	1.71	1.39-2.12	<0.005	0.56	2.26	1.68-3.05	<0.005	0.58
TMB_SYN	1.68	1.31-2.15	<0.005	0.54	2.46	1.78-3.40	<0.005	0.56
TMB_OTH	1.71	1.35-2.16	<0.005	0.55	2.43	1.78-3.32	<0.005	0.58
Multivariate								
TMB_NS	1.45	1.11-1.90	0.01	0.57	1.55	1.04-2.31	0.03	0.6
TMB_SYN	1.13	0.81-1.58	0.48		1.41	0.89-2.24	0.14	
TMB_OTH	1.26	0.92-1.74	0.16		1.48	0.94-2.34	0.09	

patterns of patients with TMB_NS less than 35 and greater than 35 were observed. For PFS, the observed p-value was 0.04, and for OS, the observed p-value was 0.022 (Figure 4.13). The patients with TMB_NS greater than 35 are hypermutators, and the characteristics specific to these high-risk patients were examined thoroughly.



(a) Overall survival

(b) Progression Free Survival

Figure 4.13: High TMB is associated with poor overall survival in NDMM patients. The difference in the overall survival probability between low and high TMB_NS is statistically significant with *p*-values 0.045 and 0.022 for PFS and OS respectively.

4.3.7 Comparison of TMB and SBS based on the overall survival event

Out of 832 MM patients for which survival data were available, 177 observed poor OS outcome while the rest of the 655 MM patients observed superior OS outcome. SBS and TMB values of the two groups were examined, and Wilcoxon rank-sum test was used to deduce if the change in the TMB and SBS values is statistically significant or not. The median SBS and TMB values for the two groups are shown in Table 4.5. There was a

significant change (p-value <0.05) for SBS T>G, C>A, and C>T. An increase was observed in the C>A and C>T substitution values, while a decrease was observed in T>G substitutions. Further, there is a statistically significant difference in the TMB values of TMB_NS, TMB_SYN, and TMB_OTH, i.e. there was a considerable increase in the tumor mutational burden of the patients with poor outcome as compared to patients with a superior outcome.

Table 4.5: The table shows the median values of TMB and SBS for the two groups of MM patients, one where the death event was observed and the other where the death event was not observed. Wilcoxon rank-sum test was applied to determine if the change is statistically significant or not. For substitutions, C>A, C>T, and T>G, the frequency was statistically different (p-values < 0.05) between the two groups.

		Median (OS event = 0)	**Median (OS event = 1)**	***p-value***
SBS	C>A	17	18	0.018
	C>G	20	21	0.1205
	C>T	59	64	0.038
	T>A	17	16	0.07
	T>C	36	33	0.08
	T>G	19	17	0.02
TMB	TMB_NS	0.4828	0.5766	4.26E-07
	TMB_SYN	0.3487	0.4023	0.002
	TMB_OTH	1.341	1.5288	3.08E-04

4.4 Discussion

The fundamental goal of the study was to investigate the entire spectrum of the mutations altered in MGUS and MM, thereby identifying the critical factors responsible for the progression of the disease from MGUS to MM. In this study, we have explored the nonsynonymous and synonymous variants due to their impact on protein expression and function. First of all, variants were identified using four different variant callers to reduce the false positives from the study. Our approach ensured that the variants discovered in our research are the closest possible estimation of the true variants present in the MM and MGUS patients. Variants were then categorized into three main categories-nonsynonymous (NS), synonymous (SYN), and other (OTH) variants. TMB was calculated for each of the three categories of variants. This study reveals changes in the mutational spectrum from MGUS to MM. There was a statistically significant rise in the single base substitutions as the disease progressed from MGUS to MM (Figure 4.2). The frequency pattern of the substitutions in MM is similar to what was observed in a previous study [63]. The highest rise in the frequency was observed in C>T transitions, where the median almost doubled from 30 to 59. An increase in the C>T transitions in

MM can be attributed to the overexpression of A3B, an APOBEC cytidine deaminase, that has an essential part in immunity against diseases [166]. Aberrant expression of A3B is known to be correlated with drug resistance, metastasis, and poor prognosis in breast cancer [167], lung cancer [168], and ovarian cancer [169]. Yamazaki et al. [166] proposed that A3B may promote disease progression and drug resistance in MM, which validates our observation of the hike in C>T transitions from MGUS to MM. The association of the frequency of substitutions in the MM patients and their survival outcome was further explored. Frequency of C>T, C>A, and T>G substitutions were significantly higher in MM patients with poor overall survival outcome as compared to MM patients with superior overall survival outcome (Table 4.5). However, in multivariate Cox Hazard analysis (Table 4.2), only C>A transitions have a statistically significant impact on the survival outcome of MM patients.

In addition, SBS2 and SBS13 mutational signatures are linked to APOBEC activity reported in MM in multiple studies [170, 171]. APOBEC signatures were found in nearly 9.63% (98/1018) of the total MM patients, while they were primarily absent in MGUS patients (present in only 1 out of 61 MGUS patients). This finding suggests that ABOPEC activity may be responsible for the molecular mechanisms driving tumor progression from MGUS to MM. The association of ABOPEC activity with overall and progression-free survival in MM was also explored. There was a statistically significant association between the ABOPEC activity and poor overall survival in MM (p-value=0.0056). The KM survival analysis validated this, which yielded significant separation (p-value=1.8e-4) in the OS curves of MM patients with and without APOBEC activity. Contrary to these findings, no significant association was found between PFS and APOBEC activity. Further, signatures SBS6, SBS14, SSB20, SBS21, SBS26 were found only in MM and are associated with defective DNA mismatch repair and microsatellite instability (MSI) as described previously. MSI has been reported in Multiple Myeloma [172]. However, its frequency is low ($\tilde{1}0\%$) [173]. MSI has been observed to be an effective indicator of response to immunotherapy in solid tumors [174], like colorectal carcinoma [175]. Therefore, it is vital to look for these signatures in MM to help identify the high-risk MM patients in need of immunotherapy.

In the present study, synonymous mutations have been examined along with nonsynonymous mutations. Though synonymous mutations do not change the amino acid sequence of the resulting protein, they have a profound influence on RNA stability, RNA folding [93] or splicing [94], translation [95], or co-translational protein folding. Hence, their role in cancer progression cannot be ignored. There are three different variants categorized under synonymous- synonymous snvs, $3'$s and $5'$UTRs. A statistically significant rise in the $3'$ (p-value=2e-16) and $5'$UTR (p-value=2.7e-7) mutations were observed from MGUS and MM. $3'$ untranslated region (UTR) are a part of mRNA containing regulatory binding sites that post-transcriptionally influence gene expres-

sion and may lead to disruption in critical pathways associated with different types of cancers. Multiple studies have demonstrated that 3′ variants are linked to the risk of developing tumor or tumor progression. Zhang et al. [176] discovered that a polymorphism detected in the IL-1α 3′ of the miRNA-122-binding site was associated with the risk of epithelial ovarian cancer. A unique variant located in the 3′ was identified in the gene PCM1, which was significantly associated with ovarian cancer [177]. Recently, Melaiu et al. [178] evaluated the significance of germline genetic variants located within the 3′-untranslated region (polymorphic 3′, i.e., p3UTR) of candidate genes involved in multiple myeloma. Their findings suggested that IL10-rs3024496 was associated with an increased risk of developing MM and worse overall survival in MM patients. They also observed that IL10-rs3024496 SNP might regulate the IL10 mRNA expression and hence, could help in the stratification of MM patients in terms of risk progression and prognosis. 5′UTR regions are a part of mRNA, which regulates the protein expression by controlling the translation initiation. Hence, single nucleotide polymorphisms (SNPs) located at 5′UTR regions may alter the protein levels by regulating the mRNA translation efficiency, thereby disturbing consequential biological pathways. The role of 5′UTR variants in multiple cancers has been explored in previous studies. A 5′UTR variant was the driving factor leading to familial breast and ovarian cancer in two independent families [179]. 5′UTR SNP in the PLA2G2A gene was associated with PC metastasis [180]. Thus, it can be concluded that 3′ and 5′ UTR mutations are more frequent in MM and drive MGUS to MM via regulatory binding sites.

TMB has become a prominent biomarker of enhanced responsiveness to immunotherapy and better outcomes. High TMB is often associated with longer survival after treatment with immune checkpoint inhibitors (ICIs) [90]. However, in non-ICI-treated patients, high TMB was associated with poor prognosis and overall survival in many cancer types [91]. Correlation of high TMB with response to targeted immunotherapies has been established in solid tumors [181, 182]. High somatic mutation and neoantigen loads have been correlated with reduced PFS in MM [183]. However, the association of TMB with overall survival is still unknown in newly diagnosed multiple myeloma (NDMM) patients. Patients with very high TMB_NS values were further analyzed to examine the relation of TMB with OS. These are known as hypermutators and are high-risk patients. Hypermutators demonstrated a significant poor overall survival (p-value=0.022) and poor progression-free survival (p-value=0.045) as compared to non-hypermutators (TMB_NS$\leq$35) (Figure 4.13). The median overall survival of hypermutators was 220 weeks compared to 316 weeks of non-hypermutators, while the median progression-free survival of hypermutators was 105 weeks compared to 143.3 weeks non-hypermutators. Mutational signatures SBS1, SBS5, and SBS54 were observed in hypermutators and death events in 7 out of 10 hypermutators. DBS4, DBS5, DBS9, DBS10, and DBS11 are the mutational signatures reflective of double base substitutions

(DBS) and were found to be present in hypermutators. On the contrary, no DBS signatures were found in low TMB patients (TMB_NS<0.1; n=6). SBS1 and SBS5 were present in low TMB patients, including SBS7a, SBS17b, SBS27, SBS51, and SBS86. Our study establishes that the frequency of hypermutators is low in the MM population, and hypermutators are associated with poor OS and poor PFS outcome. Since TMB is a predictor of enhanced responsiveness to immunotherapy, hypermutators may be treated with immunotherapy drugs such as Daratumumab/Elotuzumab [184], Isatuximab [185], and Belantamab Mafodotin [186] to improve their overall survival.

4.4.1 Conclusion

In conclusion, the present study revealed the factors responsible for disease progression from MGUS to MM and poor survival outcome in MM via a detailed investigation of the mutations present in MGUS and MM. The entire landscape of the mutational spectrum involving both synonymous and nonsynonymous mutations was examined. This study finds a change in the mutational spectrum with a statistically significant increase from MGUS to MM. There was a statistically significant increase in the frequency of all the three categories of variants-non-synonymous, synonymous, and others from MGUS to MM ($p<0.05$). However, there was a statistically significant rise in the TMB values for TMB_NS and TMB_SYN only. We also observed that $3'$ and $5'$ UTR mutations were more frequent in MM and might be responsible for driving MGUS to MM via regulatory binding sites. A detailed investigation of these mutations might help understand the mechanism of the progression of MGUS to internecine MM and may be explored in future studies. In addition, NDMM patients were also examined separately along with their survival outcome. 10 out of 832 NDMM patients had TMB_NS values greater than 35 and were designated as hypermutators. It could be concluded that the frequency of hypermutators was low in MM with poor OS and PFS outcome. We also observed a statistically significant rise in the frequency of C>A and C>T substitutions and a statistically significant decline in T>G substitutions. There was a statistically significant increase in the tumor mutational burden of the patients with poor outcome as compared to patients with a superior outcome. Further, a statistically significant association between the APOBEC activity and poor overall survival in MM was discovered. A limitation of the current study is that the number of MGUS patients is significantly less than the number of MM patients. Comparison with a larger cohort of MGUS patients can substantiate the findings of the study. A coherent analysis of evolving mutational landscapes and cancer signatures could assist in designing therapies to impede the transformation of benign MGUS to malignant MM. Additionally, a systematized comparison of high-risk MM patients with low-risk MM patients can aid in identifying the risk factors responsible for disease progression and ultimately guide towards a per-

sonalized cure, thereby improving the overall survival of MM patients. Assessing the risk stage of MM patients based on their genomic profile is a challenging field, since, it is dependent on many prognostic factors. This motivated us to our next research problem of developing a robust system for risk stratification in MM. The proposed method has been explained in detail in chapter 5.

Chapter 5

AI-supported risk staging system for multiple myeloma

5.1 Introduction

The main objective of risk staging system is to identify high risk patients in order to optimize their treatment and improve their overall survival. Moreover, it also aids in enhancing the positive outcomes in low risk and medium risk patients. Risk staging systems are mostly developed using the genetic and clinical features. However as discussed in section 1.2, ethnicity plays an important role in disease biology and must be considered while designing an optimal risk staging model. Therefore, to address this concern, we first investigated the role of ethnicity in differential clinical characteristics between the two independent cohorts of MMIn (MM patients belonging to the Indian population) and MMRF (Multiple Myeloma Research Foundation) in this study. Both these datasets belong to patients with newly diagnosed multiple myeloma (NDMM) belonging to two separate ethnic groups. Further, we proposed a Consensus based risk-stratification system (CRSS), an AI-enabled risk-stratification system, for NDMM that incorporates the ethnicity-specific cut-offs of the laboratory parameters like albumin, beta-2 microglobulin (β2M), calcium, estimated glomerular filtration rate (eGFR), hemoglobin, age along with high risk cytogenetic abnormalities (HRCA). The newly proposed ethnicity aware AI-assisted CRSS method was shown to have superior performance as compared to R-ISS. In addition, we also interpreted our proposed model via SHAP [187] analysis to demonstrate the clinical significance of the risk stage predictions by CRSS. Our findings establish the significance of integrating ethnicity-specific information as well as the effectiveness of machine learning methods in devising a robust risk-staging model for MM.

5.2 Materials and Methods

5.2.1 Datasets

A total of 1675 entries were found in the computerized database search on June 28, 2019 with keyword 'ICD C90' registered at the Institute Rotary Cancer Centre, All India Institute of Medical Sciences (AIIMS). Patients with plasma cell dyscrasia other than MM (n=253) or who were lost to follow up after a single visit (n=111) or before

first response could be assessed (n=21) or with inadequate clinical and/or laboratory parameters (n=121) or with early deaths (n=99) were excluded. Remaining 1070 patients of MM belonging to the Indian population, referred to as MMIn, were evaluated in this study (Figure 5.1). Out of 1070 patients, 41 patients had one or two missing values. There are several methods to impute missing values [188, 189, 190, 191]. However, in the MMIn dataset, missing values were imputed with the median value of the parameters. An independent cohort of 900 MM patients enrolled in Multiple Myeloma Research Foundation (MMRF) repository, was also used for developing the model. Clinical and laboratory data for MMRF dataset, belonging to the American population, is available publicly. High risk cytogenetic information was available for 384 out of 1070 patients in the MMIn cohort and 800 out of a total of 900 patients in the MMRF which were further used for building the staging model.

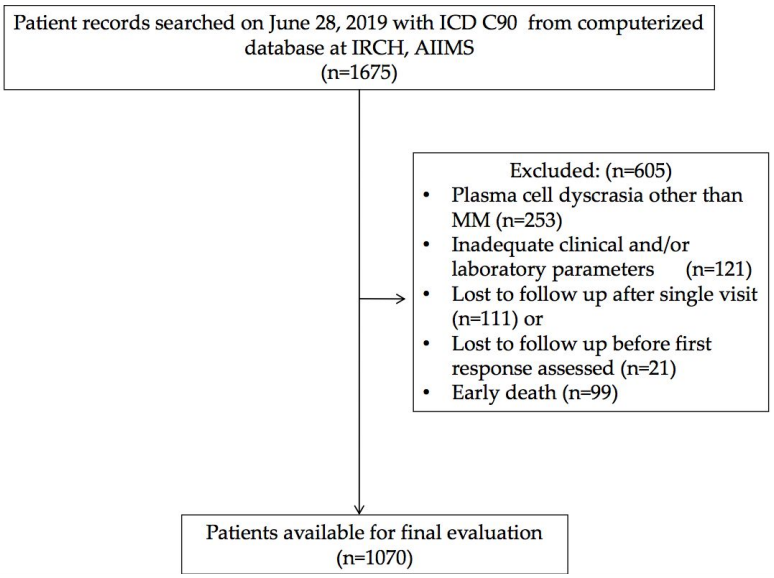


Figure 5.1: Flowchart of Study Population for MMIn dataset.

5.2.2 Clinical and Laboratory Characteristics

The clinical, laboratory, and radiological data was obtained from the medical case files. R-ISS could be assigned to a subset of patients (n=627) as described previously [192]. Response outcome was estimated following the International uniform response criteria for multiple myeloma [193]. Progression free survival (PFS) was computed from the date of diagnosis till the time of progression or death. Overall survival (OS) was computed from the date of diagnosis till death due to any cause or being censored at last follow-up. Baseline clinical and laboratory features of the patients are given in the Table 5.1.

Table 5.1: Baseline demographic, laboratory and clinical characteristics of multiple myeloma (MM) patients of MMIn and MMRF cohort.

Parameters	**MMIn (*n*=1070)**	**MMRF (*n*=900)**
Age (Median, Range; in years)	56 (18-87)	62 (27 - 91)
Male/ Female	710 (66·36%) 360 (33·64%)	529 (58.78%) 371 (41.22%)
Hemoglobin (g/dL) <10 $\geq$10	599 (55·98%) 471 (44·02%)	331 (36·77%) 569 (63·23%)
Serum albumin (g/dL) <3·5 $\geq$3·5	449 (41·96%) 621 (58.04%)	328 (36·44%) 572 (63·56%)
Beta 2 microglobulin (mg/L) <5·5 $\geq$5·5	534 (49.90%) 536 (50.09%)	661 (73·44%) 239 (26·56%)
Serum LDH (IU/L) $\leq$280 >280	929 (86·82%) 141 (13·18%)	850 (94·44%) 50 (5·56%)
Serum creatinine (mg/dL) $\leq$2 >2	830 (77.57%) 240 (22.43%)	816 (90·66%) 84 (9·34%)
Serum calcium (mg/dL) $\leq$11 >11	935 (87.38%) 135 (12.62%)	831 (92·33%) 69 (7·67%)
ISS 1/2/3	207/323/540	342/319/239
R-ISS 1/2/3	47/459/121	107/505/91

5.2.3 Study Design

The complete design strategy of the Consensus based approach for developing the risk-stratification system (CRSS) is explained in this section (Figure 5.2). Data from both the cohorts was separately used to develop the risk-staging models based on CRSS. Different clinical parameters were evaluated for developing the risk staging system consisting of age, albumin, β2M, calcium, eGFR, Hemoglobin, LDH and HRCA which includes t(4;14), t(14;16) and del17. β2M and LDH levels are reflective of tumor burden and serum albumin, hemoglobin, calcium and creatinine are reflective of the bone and renal homeostasis. eGFR was calculated from creatinine concentration using MDRD eGFR equation [194]. LDH values were brought to a common scale by multiplying each entry by 280 and dividing it by the upper limit of LDH provided for that particular entry in MMIn data. Description of the steps used in consensus based approach for developing risk staging model is given below:

Step 1: Dividing patients into two risk groups based on established thresholds of parameters: For each parameter, patients were initially divided into high-risk and low-risk groups using the well-established cut-offs of these parameters as shown in Table 5.2.

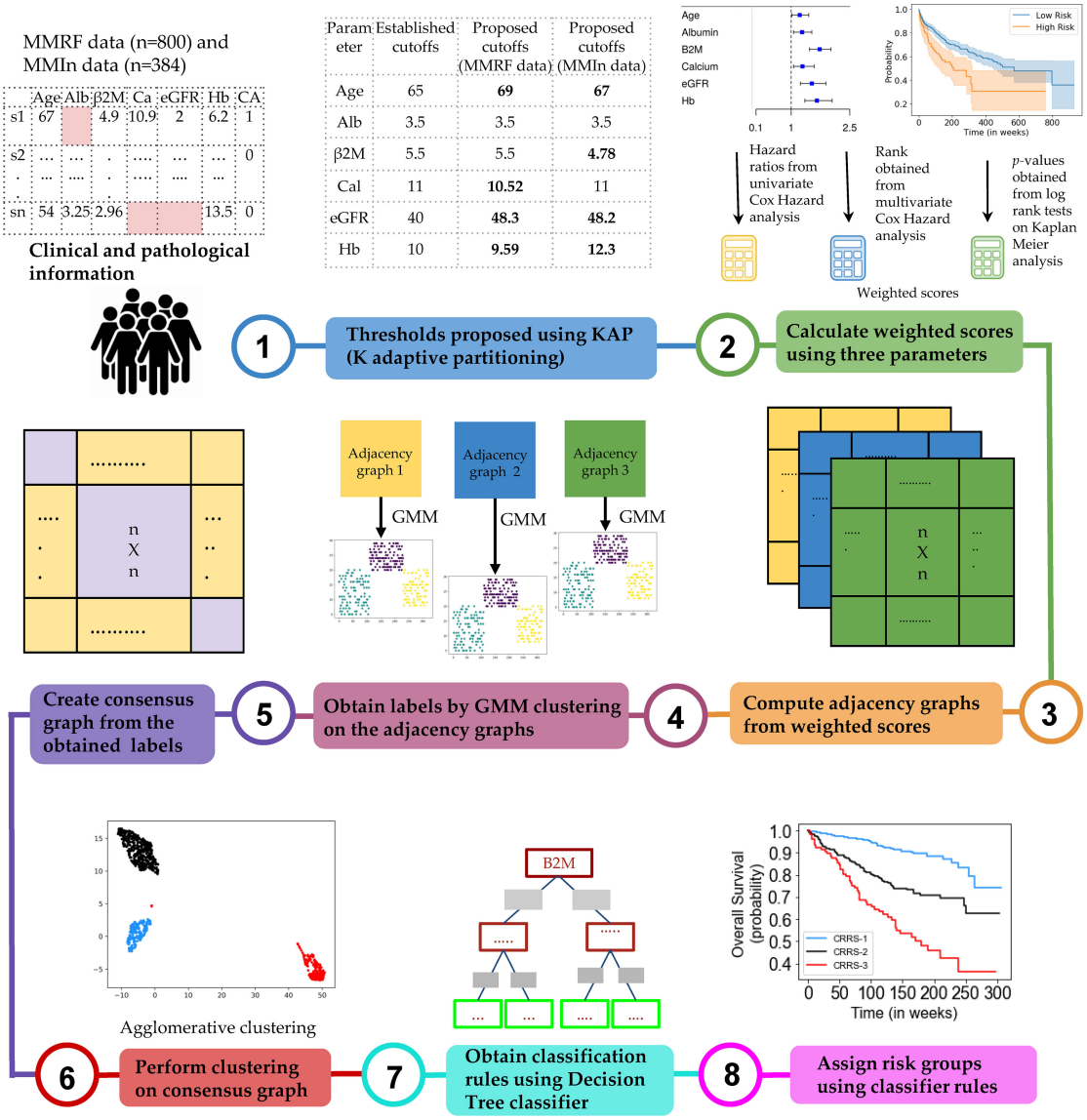


Figure 5.2: Comparison of established and proposed cutoffs for clinical and laboratory parameters for the stratification of patients for progression-free survival (PFS) and overall survival (OS) in MMIn and MMRF using Kaplan–Meier analysis.

Established thresholds for albumin and β2M are derived from ISS and for eGFR, calcium, hemoglobin are derived from revised IMWG criteria [195].

Step 2: Finding new thresholds of parameters via KAP: K-adaptive partitioning [164] (KAP) algorithm was used to find new threshold values for the parameters using complete data of MMIn (n=1070) and MMRF (n=900). KAP was performed on the patients' parameters yielding two threshold values for each parameter, one from PFS and the other from OS analysis. The cut-off which was close to the original value was chosen as the new cut-off for each parameter. Patients were again divided into high and low risk groups based on the proposed cut-offs. Proposed thresholds maximised the

Table 5.2: Comparison of established and proposed cutoffs for clinical and laboratory parameters for the stratification of patients for progression-free survival (PFS) and overall survival (OS) in MMIn and MMRF using Kaplan–Meier analysis. Note: The proposed cutoffs were found using complete data of MMIn (n = 1070) and MMRF (n = 900). Less than or equal to cutoff reveals the increased risk in the patient. ">65" shows that a patient with age greater than 65 years is at greater risk than a patient less than 65 years. "$\leq$3.5" shows that a patient with albumin levels less than equal to 3.5 is at a greater risk than a patient with albumin levels greater than 3.5. It holds true for other parameters also in a similar manner. Bold values of the column "proposed cutoff value" signify the change in the value of the parameters from the existing cutoffs. p-values in bold signify that p-values became more significant with the proposed changes in cutoffs.

Parameter	**Established cutoff value**	**Proposed cutoff value**	**PFS**		**OS**	
			***p*-value with established cutoff**	***p*-value with proposed cutoff**	***p*-value with established cutoff**	***p*-value with proposed cutoff**
MMIn (n = 1070)						
Age (years)	>65	**>67**	0.11	**0.012**	5.84e-5	**1.25e-6**
Albumin (g/dl)	$\leq$3.5	$\leq$3.5	0.115	0.115	7.0e-4	7.0e-4
β2M (mg/L)	$\geq$5.5	$\geq$**4.78**	**8.15e-10**	9.32e-10	4.13e-10	**4.53e-14**
Calcium (mg/dl)	$\geq$11	$\geq$11	0.0078	0.0078	0.0037	0.0037
eGFR (ml/min/1.73m^2)	$\leq$40	$\leq$**48.2**	0.16	**0.04**	0.005	**1.5e-4**
Hb (g/dl)	$\leq$10	$\leq$**12.3**	0.0019	**8.56e-5**	0.0014	**3.75e-7**
MMRF (n = 900)						
Age (years)	>65	**>69**	3.23e-05	**1.98e-08**	1.06e-05	**1.58e-09**
Albumin (g/dl)	$\leq$3.5	$\leq$3.5	0.00017	0.00017	8.47e-07	8.47e-07
β2M (mg/L)	$\geq$5.5	$\geq$5.5	1.22e-10	1.22e-10	9.25e-13	9.25e-13
Calcium (mg/dl)	$\geq$11	$\geq$**10.52**	0.0077	**1.40e-04**	5.88e-06	**3.49e-06**
eGFR (ml/min/1.73m^2)	$\leq$40	$\leq$**48.3**	4.5e-05	**4.67e-09**	7.48e-06	**2.48e-10**
Hb (g/dl)	$\leq$10	$\leq$**9.59**	2.82e-06	**5.69e-09**	6.77e-06	**5.42e-07**

separation between the high and low risk groups as compared to the established thresholds. This is evident from the lower p-values obtained from the Log-rank test on the Kaplan-Meier curves for all the parameters. A complete list of proposed thresholds for MMIn and MMRF data is shown in Table 5.2.

Step 3: Cumulative integration of the prognostic impact of the parameters: The collective prognostic impact of the parameters was integrated into risk staging via creation of

three different adjacency graphs using hazard ratios obtained from univariate Cox hazard analysis, p-values obtained from Log Rank test on Kaplan-Meier curves and ranks obtained from multivariate Cox hazard analysis.

Step 4: Creation of first adjacency graph: First adjacency graph was created using ranks obtained from the multivariate Cox hazard analysis. The parameter with the highest hazard value was given the highest rank and the one with the lowest hazard value was given the lowest rank. The respective ranks served as the weights of each of the parameters and captured the relative impact of each parameter on the patients' survival. Next, the risk score for each patient was calculated by successive addition of the weights of all those parameters that had values (in the respective patient) greater than the cut-offs defined for the high-risk group. These patient scores were used to compute an adjacency graph of *n* rows and *n* columns (columns are features) where n is the number of patients. Each row corresponds to one patient and each entry in the row is the absolute difference between the score of that patient with each of the patients including self.

Step 5: Creation of second and third adjacency graphs: For the second adjacency graph, hazard ratio values obtained from univariate Cox hazard analysis were used. For each parameter, the highest of the two HR values obtained from PFS and OS was chosen and normalized using 'minmax' scaling. The scaled HR values were assigned as the respective weights of each of the parameters representing the impact of each parameter on patients' survival. Third adjacency graph was created using p-values obtained by performing a Log-rank test on Kaplan-Meier curves. For each parameter, the lower of the two p-values obtained from PFS and OS was chosen and normalized using 'minmax' scaling. The scaled p-values were assigned as the respective weights of each of the parameters. Further, the risk score for each patient was calculated by successive addition of the weights of all those parameters that had values (in the respective patient) greater than the cut-off defined for the high-risk group. The two different patient scores obtained from univariate hazard ratios and p-values were further used to compute two separate adjacency graphs of n rows and n columns (columns are features) where n is the number of patients. Each row corresponds to one patient and each entry in the row is the absolute difference between the score of that patient with each of the patients including self.

Step 6: GMM clustering on the adjacency graphs: Gaussian mixture model (GMM) based clustering is an unsupervised clustering algorithm which was applied on the three adjacency graphs to obtain clustering labels.

Step 7: Creation of a consensus graph: The clustering outputs of the three different adjacency graphs were used to create a consensus graph [196] of size $n \times n$. The entry for the i^{th} row and j^{th} column in the consensus graph was determined by calculating the number of times i^{th} and j^{th} patients were assigned the same group. Diagonal entries

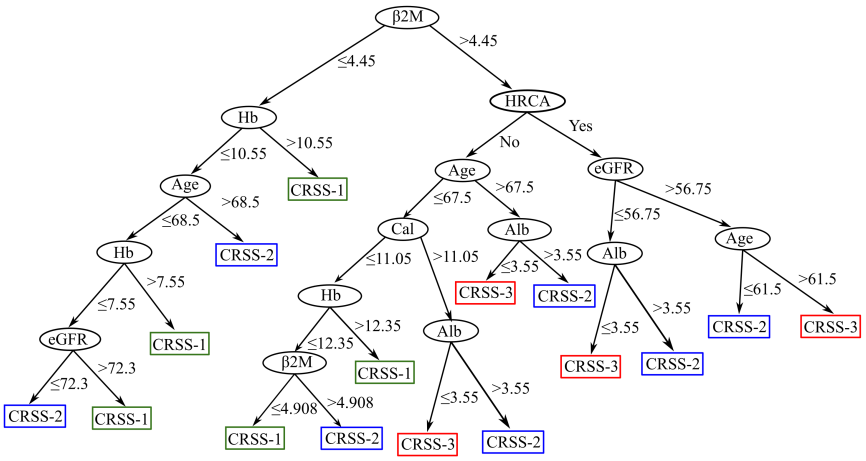
were zero in this graph.

Step 8: Hierarchical clustering on the consensus graph: Agglomerative clustering was performed on the consensus graph to cluster the patients into three risk groups. Each cluster of patients was assigned one label: Stage-1 (low-risk), Stage-2 (intermediate-risk), or Stage-3 (high-risk). The rationale behind using multiple clustering was to combine the results of the clustering outputs achieved from the different adjacency graphs and ensure the stability of the final clusters so deduced from agglomerative clustering.

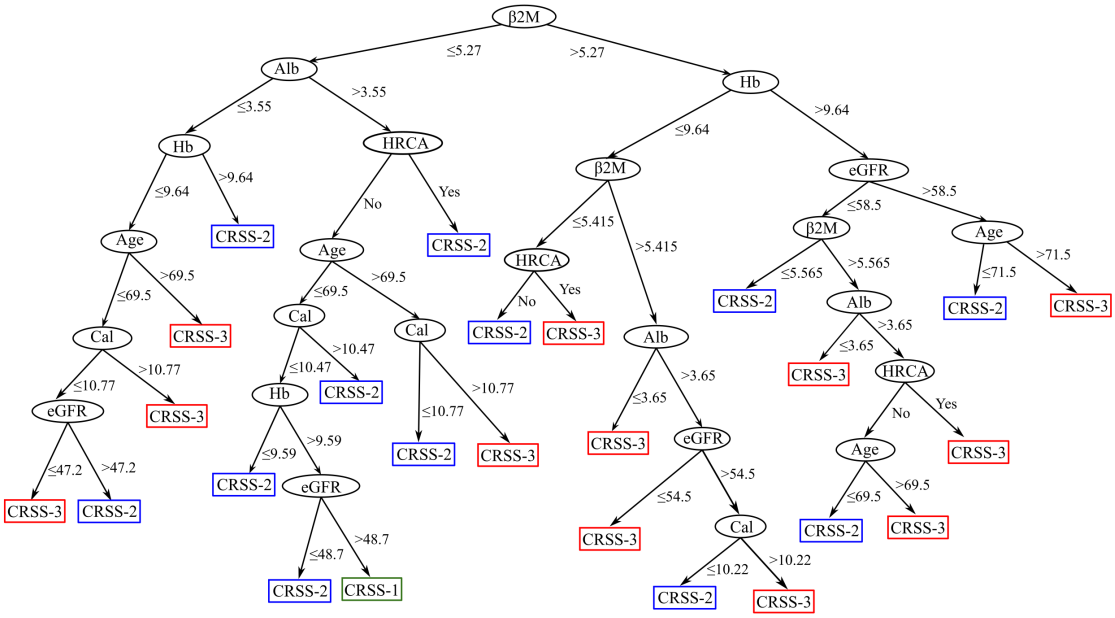
Step 9: Training a Decision tree classifier: The staging labels obtained from agglomerative clustering served as ground truth labels for training the supervised Decision tree classifier. The trained Decision tree classifier provided the rules in terms of the parameters for the identification of risk groups, labeled as CRSS-1 (low risk), CRSS-2 (intermediate-risk), and CRSS-3 (high-risk) (Figure 5.3). Step 10: Infer actual risk groups of the patients using Decision tree classifier rules: Decision tree classifier rules were then used to identify the risk stages of the patients in both the cohorts. The risk stage assigned by the Decision tree classifier was considered the actual risk class for each patient.

5.2.4 Creation of multiple models on the datasets

CRSS method explained in Figure 5.2 was used to create multiple models for MMIn and MMRF datasets. Models A1, A2 and A3 were built for MMIn data (Table 5.3). Model A1 was built using established cut-offs of the parameters of albumin, β2M, LDH and HRCA. Model A2 was built using the established cut-offs of the parameters of albumin, age, calcium, eGFR, hemoglobin, β2M and HRCA. Model A3 uses the same parameters as Model A2, but with the newly proposed cut-offs of the parameters derived from MMIn dataset. Similarly, models M1, M2, M3 and M4 were built for MMRF data (Table 5.3). Model M1 and M2 are equivalent to Model A1 and A2 respectively. For the model M3, proposed cut-offs of parameters derived from MMIn dataset were used for albumin, age, calcium, eGFR, hemoglobin, β2M and HRCA. Model M4 is similar to Model M3, but uses the proposed cut-offs of the parameters derived from the MMRF dataset.



(a)



(b)

Figure 5.3: Hierarchical rule based tree structure to assign data samples to CRSS-1 (Low), CRSS-2 (Inter) and CRSS-3 (High) groups. Parameters: Age: Age; Alb: Albumin; β2M: beta2-microglobulin; Ca: Calcium; eGFR: estimated glomerular filtration rate; Hb: hemoglobin and HRCA: High risk cytogenetic abnormalities. (a) MMIn cohort and (b) MMRF cohort

Table 5.3: Comparison of different models devised for the risk stratification of patients in the MMIn and MMRF cohorts with the R-ISS. Models were built using data for which high-risk cytogenetic information (HRCA) was available (n = 384 for MMIn and n = 800 for MMRF). R-ISS information was available for only 355 out of 384 patients in the MMIn dataset and 658 out of 800 patients in the MMRF dataset. The model with the best performance was A3 and M4 (in bold). Model A1: beta-2 microglobulin (β2M), albumin, LDH, and CA [del17, t(4;14), t(14;16)] at existing cutoffs. Model A2: age, β2M, albumin, calcium, estimated glomerular filtration rate (eGFR), Hb, and HRCA using existing cutoffs. Model A3: age, β2M, albumin, calcium, eGFR, Hb, and HRCA using proposed cutoffs for MMIn data. Model M1: β2M, albumin, LDH, and HRCA at existing cutoffs. Model M2: age, β2M, albumin, calcium, eGFR, Hb, and HRCA using existing cutoffs. Model M3: age, β2M, albumin, calcium, eGFR, Hb, and HRCA using proposed cutoffs for MMIn data. Model M4: age, β2M, albumin, calcium, eGFR, Hb, and HRCA using proposed cutoffs for MMRF data.

		PFS			**OS**		
		Hazard ratio	***p*-value**	**C-index**	**Hazard ratio**	***p*-value**	**C-index**
MMIn (*n*=384)							
R-ISS (*n*=355)		1.42	0.004	0.57	2.32	<5e-6	0.636
	2vs1	1.24	0.33		2.31	0.04	
	3vs1	1.92	0.009		5.37	0.00013	
Model A1		1.5	1.00e-5	0.594	2.03	<5e-6	0.646
	2vs1	1.53	0.007		2.13	0.0013	
	3vs1	2.26	2.00e-5		4.16	<5e-6	
Model A2		1.4	0.0001	0.579	1.74	1.00e-5	0.616
	2vs1	1.42	0.056		1.9	0.02	
	3vs1	1.98	0.00013		3.13	2.00e-5	
Model A3 (CRSS)		**1.8**	<5e-6	**0.6**	**2.43**	<5e-6	**0.67**
	2vs1	**1.76**	3.00e-4		**3.95**	<5e-6	
	3vs1	**3.27**	<5e-6		**6.43**	<5e-6	
MMRF (*n*=800)							
R-ISS (*n*=658)		1.61	0.00001	0.578	2.26	<5e-6	0.618
	2vs1	1.49	0.015		1.79	0.03	
	3vs1	2.6	0.00001		4.66	<5e-6	
Model M1		1.55	<5e-6	0.6	2.07	<5e-6	0.656
	2vs1	1.55	0.00042		2.06	0.00067	
	3vs1	2.4	<5e-6		4.3	<5e-6	
Model M2		1.62	<5e-6	0.6	2.36	<5e-6	0.657
	2vs1	1.44	0.01		2.12	0.0081	
	3vs1	2.54	<5e-6		5.22	<5e-6	
Model M3		1.54	<5e-6	0.604	2.2	<5e-6	0.679
	2vs1	1.87	<5e-6		2.95	<5e-6	
	3vs1	2.32	<5e-6		5.11	<5e-6	
Model M4 (CRSS)		**1.79**	<5e-6	**0.61**	**2.85**	<5e-6	**0.676**
	2vs1	**1.76**	8.10e-4		**4.1**	3.40e-4	
	3vs1	**3.19**	<5e-6		**10.61**	<5e-6	

5.3 Results

5.3.1 Clinical and Laboratory characteristics of myeloma patients

The baseline clinical and laboratory features of patients from the two cohorts were compared using unpaired Wilcoxon rank-sum test. The median values of all the parameters

except albumin was found to be significantly different (p-value < 0.05, Table 5.4) in both the cohorts thereby substantiating that the two populations are different. Novel agents (IMIDs:thalidomide or lenalidomide and/or PSI i.e. bortezomib) either as primary or maintenance therapy were given to all the patients. Triplet Therapy was rendered to 56·5% of patients. With a median follow up of 166 weeks (range: 14-961 weeks), 626 patients progressed (median PFS =117 weeks) and 372 died (median OS=166 weeks).

Table 5.4: The parameters of the two cohorts MMIn and MMRF were compared via unpaired Wilcoxon rank-sum test. If the *p*-value < 0.05, it can be concluded that the median is significantly different in both the cohorts. Median value of albumin was not statistically different between MMIn and MMRF, while these were statistically different for the rest of the parameters across the cohorts.

Parameter	***p-value***
Age	3.09e-34
Albumin	0.2
β2M	2.54e-34
calcium	0.00029
eGFR	1.98e-09
Hemoglobin	2.89e-34

5.3.2 Results on MMIn dataset (*n*=384)

Univariate Cox analysis of the entire patient cohort (n=1070, Table 5.5, Figure 5.4), revealed increased risk of progression and mortality for age>67 years, albumin$\leq$3.5, β2M$\geq$4.78, calcium$\geq$11, eGFR$\leq$48.2 and hemoglobin$\leq$12.3. Multivariate Cox hazard analysis was also performed to analyse the cumulative risk of the parameters (Table 5.6). Of the three models generated, model A3 based on ML derived cut-offs for the prognostic parameters was best with higher C-index and hazard ratio (Table 5.3). Using model A3, the patients were risk stratified and the largest proportion of patients were placed in CRSS-2 (n=192, 50%) followed by CRSS-1 (n=137, 35.68%) and CRSS-3 (n=55, 14.32%). KM survival analysis of CRSS groups indicated statistically significant difference in PFS between CRSS-1 and CRSS-2 groups (median PFS: 213 vs. 138 weeks; p=0.0003) and between CRSS-2 and CRSS-3 groups (median PFS: 138 vs.100 weeks; p=0.0026) (Figure 5.4). For R-ISS, there was a statistically significant difference in PFS between R-ISS2 and R-ISS3 (median PFS: 160 vs.105 weeks; p=0.01) but not between R-ISS1 and R-ISS2 (median PFS=196 vs. 160 weeks p=0.31). Further, for CRSS there was statistically significant difference in OS between CRSS-1 and CRSS-2 groups (median OS= 495 vs. 249 weeks; p=1.08e-8) as well as between CRSS-2 and CRSS-3 groups (median OS=249 vs. 182 weeks; p=0.02). For R-ISS, there was statistical difference in OS between R-ISS2 and R-ISS3 groups (median OS=377 vs.

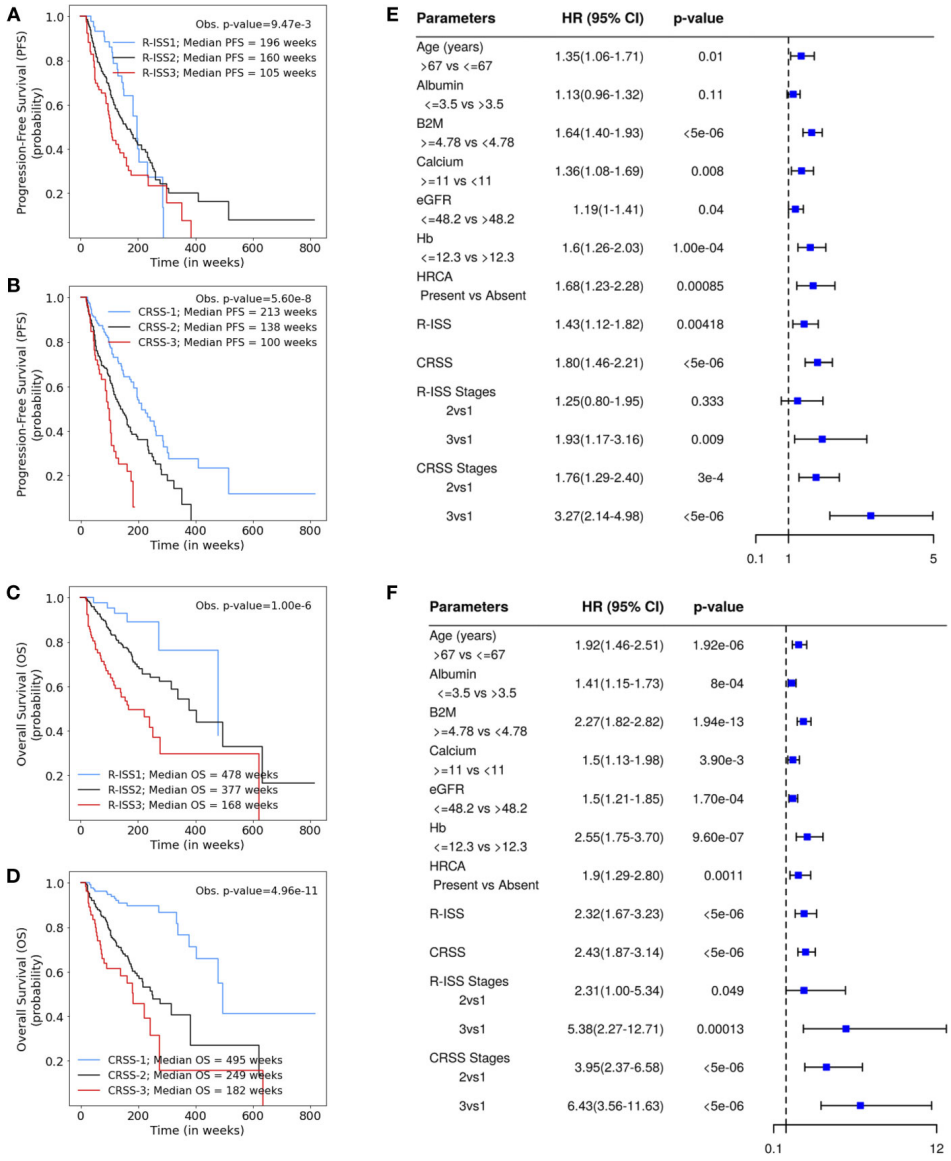


Figure 5.4: (A, B) Progression-Free Survival (PFS) in patients with multiple myeloma (MM) from the MMIn cohort (n = 1070) stratified by the R-ISS (n = 355) and the proposed CRSS (n = 384), respectively. Median PFS for R-ISS1, R-ISS2, and R-ISS3 are 196, 160, and 105 weeks, respectively. Observed p-value obtained after performing a log-rank test on R-ISS is 9.47e-3. Median PFS for CRSS-1, CRSS-2, and CRSS-3 are 213, 138, and 100 weeks, respectively. Observed p-value obtained after performing a log-rank test on CRSS is 5.60e-8. (C, D) Overall survival (OS) in patients with MM from the MMIn cohort (n = 1070) stratified by the R-ISS (n = 355) and CRSS (n = 384), respectively. Median OS for R-ISS1, R-ISS2, and R-ISS3 are 478, 337, and 168 weeks, respectively. Observed p-value obtained on R-ISS is 1.00e-6. Median OS for CRSS-1, CRSS-2, and CRSS-3 are 495, 249, and 182 weeks, respectively. Observed p-value obtained on CRSS is 4.96e-11. (E, F) Univariate Cox hazard analysis on the prognostic factors. Hazard ratios for all the parameters except HRCA were calculated on complete data (n = 1070) for the MMIn dataset. Hazard ratio for HRCA and the risk-staging models were found using the data for which HRCA information was present (n = 384 for the MMIn dataset).

168 weeks; p=1.86e-5) as well as between R-ISS1 and R-ISS2 groups (median OS=478 vs. 377 weeks; p=0.03).

Table 5.5: Univariate Cox hazard analysis on the prognostic factors- age, albumin, β2M, calcium, eGFR, hemoglobin and high risk cytogenetic abnormalities (HRCA). Hazard ratios of all the parameters except HRCA were calculated on the full data (n=1070 for MMIn and n=900 for MMRF). Hazard ratio of HRCA was found using data for which HRCA information was present (n=384 for MMIn and n=800 for MMRF).

Parameter (lower risk threshold, higher risk threshold)	**MMIn (n=1070, HRCA available for n=384)**					
	PFS			**OS**		
	HR	**CI**	***p-value***	**HR**	**CI**	***p-value***
Age ($\leq$67, >67)	1.35	1.06-171	0.01	1.92	1.46-2.51	1.92e-06
Albumin (>3·5, $\leq$3·5)	1.13	0.96-1.32	0.11	1.41	1.15-1.73	8e-04
β2M (<4·78, $\geq$ 4·78)	1.64	1.4-1.93	1.34e-09	2.27	1.82-2.82	1.94e-13
Calcium (<11, $\geq$ 11)	1.36	1.08-1.69	0.008	1.50	1.13-1.98	3.9e-3
eGFR (>48·2,$\leq$48·2)	1.19	1.00-1.41	0.04	1.50	1.21-1.85	1.7e-04
Hb (>12·3, $\leq$ 12·3)	1.6	1.26-2.03	1.00e-04	2.55	1.75-3.7	9.60e-07
HRCA (del17, t(4;14), t(14;16))	1.68	1.23-2.28	0.00085	1.9	1.29-2.8	0.00112
Parameter (lower risk threshold, higher risk threshold)	**MMRF (n=900, HRCA available for n=800)**					
	PFS			**OS**		
	HR	**CI**	***p-value***	**HR**	**CI**	***p-value***
Age ($\leq$69, >69)	1.79	1.45-2.20	<5e-06	2.41	1.79-3.23	<5e-06
Albumin (>3·5, $\leq$ 3·5)	1.44	1.19-1.75	0.0002	2.06	1.53-2.76	<5e-06
β2M (<5·5, $\geq$ 5.5)	1.92	1.56-2.35	<5e-05	2.76	2.06-3.69	<5e-06
Calcium (<10.52, $\geq$ 10.52)	1.67	1.28-2.19	0.00017	2.24	1.58-3.18	1e-05
eGFR (>48·3,$\leq$48·3)	1.91	1.53-2.38	<5e-05	2.57	1.90-3.49	<5e-06
Hb (>9.59, $\leq$ 9.59)	1.80	1.47-2.20	<5e-05	2.07	1.55-2.78	<5e-06
HRCA (del17, t(4;14), t(14;16))	1.08	0.87-1.35	0.48012	1.38	0.99-1.91	0.05388

C-statistic and hazard ratios computed on CRSS surpassed the C-index and hazard ratios obtained for R-ISS with respect to both PFS and OS (Table 5.3). C-statistic for CRSS

was 0.60 (AIC=2171.49, BIC=2175.43, HR=1.80, 95% CI=1.46–2.21, p<5e-6) for PFS and 0.67 (AIC=1244.72, BIC=1248.67, HR=2.43, 95% CI=1.87–3.14, p<5e-6) for OS while C-statistic for R-ISS was 0.57 (AIC=2011.14, BIC=2015.01, HR=1.43, 95% CI=1.12–1.82, p=4.18e-3) for PFS and 0.636 (AIC=1132.20, BIC=1136.07, HR=2.32, 95% CI=1.67–3.23, p < 5e-6) for OS.

5.3.3 Results on MMRF dataset (*n*=800)

For MMRF data, out of the four models generated, the model M4 performed the best and had the highest C-index and hazard ratios as compared to other models as well as R-ISS (Table 5.3). In the univariate Cox hazard analysis of the MMRF data, risk of progression and mortality was increased for age>69 years, β2M$\geq$5.5, albumin$\leq$3.5, hemoglobin$\leq$9.59, eGFR$\leq$48.3 and calcium$\geq$10.52 (Table 5.5, Figure 5.5). Multivariate Cox hazard analysis was also performed (Table 5.6). In the MMRF cohort, using the M4 model, the majority of the patients were placed in CRSS-2 (n=452, 56.5%) followed by CRSS-3 (n=174, 21.75%) and CRSS-1 (n=174, 21.75%). Results of the median PFS on CRSS groups (p=8.64e-12) and R-ISS groups (p=1.73e-5) as well as median OS on CRSS groups (p=1.08e-15) and R-ISS groups (p=6.57e-8) reveal superior performance of CRSS than R-ISS (significant p-values; Figure 5.5).

C-statistic for CRSS in MMRF data is 0.61 (AIC=4126.07, BIC=4130.74, HR=1.79, 95% CI=1.52–2.12, p<5e-6) for PFS and 0.676 (AIC=1819.95, BIC=1824.62, HR=2.85, 95% CI=2.19–3.71, p<5e-6) for OS. C-statistic for R-ISS is 0.578 (AIC=3413.36, BIC=3416.49, HR=1.61, 95% CI=1.30–2.00, p=1.00e-5) for PFS and 0.618 (AIC=1586.78, BIC=1591.27, HR=2.26, 95% CI=1.65–3.11, p < 5e-6) for OS (Table 5.3).

The 5-year OS for the MMIn (*n*=384) was 89.79% for CRSS-1, 47.91% for CRSS-2 and 31.36% for CRSS-3 (Table 5.7). Overall there is a substantial difference in the percentages of the 5-year OS and median OS for different risk groups which indicate that the groups were significant. A similar stratification was achieved when the CRSS model was applied on the MMRF test dataset. The 5-year OS for MMRF data was 94.78% for CRSS-1, 65.69% for CRSS-2 and 46.91% for CRSS-3 which is quite comparable to that obtained in the MMIn data. Higher values of C-index and hazard ratios as well as lower values of partial AIC and BIC on both the datasets were indicative of the superior performance of our AI-based CRSS method as compared to R-ISS.

Table 5.6: Multivariate Cox hazard analysis on the prognostic factors- age, albumin, β2M, calcium, eGFR, hemoglobin and high risk cytogenetic abnormalities (HRCA). Multivariate analysis was performed on data with HRCA information (n=384 for MMIn and n=800 for MMRF).

Parameter (lower risk threshold, higher risk threshold)	**MMIn (n = 384)**					
	PFS			**OS**		
	HR	**CI**	***p*-value**	**HR**	**CI**	***p*-value**
Age ($67\leq$, >67)	1.40	0.91-2.16	0.12657	2.63	1.67-4.15	0.00003
Albumin ($>3{\cdot}5$, $\leq 3{\cdot}5$)	0.92	0.70-1.22	0.57215	0.96	0.67-1.39	0.83982
β2M ($<4{\cdot}78$, $\geq 4{\cdot}78$)	1.57	1.14-2.15	0.00544	3.30	2.06-5.29	<5e-06
Calcium (<11, ≥ 11)	1.68	1.09-2.59	0.01841	1.34	0.72-2.48	0.35021
eGFR ($>48{\cdot}2$,$\leq 48{\cdot}2$)	0.91	0.66-1.25	0.56159	0.74	0.50-1.11	0.15055
Hb ($>12{\cdot}3$, $\leq 12{\cdot}3$)	1.63	0.97-2.74	0.06395	1.84	0.82-4.11	0.14009
HRCA (del17, t(4;14), t(14;16))	1.48	1.08-2.03	0.01396	1.44	0.97-2.14	0.0739
Parameter (lower risk threshold, higher risk threshold)	**MMRF (n = 800)**					
	PFS			**OS**		
	HR	**CI**	***p*-value**	**HR**	**CI**	***p*-value**
Age (≤ 69, >69)	1.52	1.20-1.92	0.00047	1.98	1.42-2.77	0.00006
Albumin ($>3{\cdot}5$, $\leq 3{\cdot}5$)	1.23	0.98-1.54	0.06812	1.74	1.23-2.45	0.00179
β2M ($<5{\cdot}5$, ≥ 5.5)	1.25	0.94-1.65	0.12029	1.48	1.00-2.20	0.04926
Calcium (<10.52, ≥ 10.52)	1.62	1.21-2.18	0.00136	1.94	1.29-2.90	0.00143
eGFR ($>48{\cdot}3$,$\leq 48{\cdot}3$)	1.19	0.89-1.60	0.24308	1.47	0.98-2.21	0.0645
Hb (>9.59, ≤ 9.59)	1.50	1.17-1.93	0.00134	1.35	0.93-1.96	0.1097
HRCA (del17, t(4;14), t(14;16))	1.11	0.89-1.39	0.34433	1.42	1.02-1.97	0.03786

5.3.4 Statistical Analysis on the parameters used in CRSS

Kruskal Wallis test was performed to compare the median values of the parameters- age, albumin, β2M, calcium, eGFR and hemoglobin across the three risk groups for both MMIn and MMRF dataset. There was a significant increase ($p < 0.05$) in the values of age and β2M while there was a significant decrease ($p<0.05$) in the values

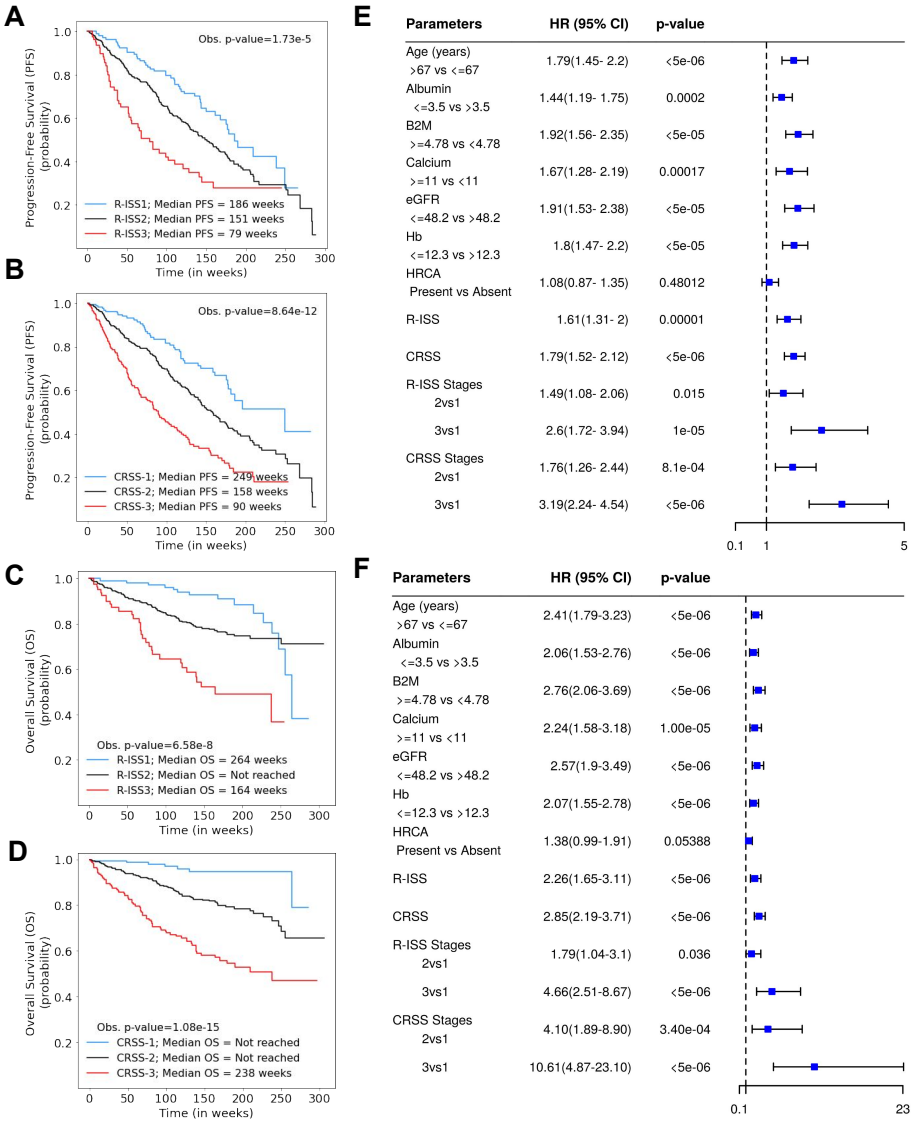


Figure 5.5: (A, B) Progression-Free Survival (PFS) in patients with MM from MMRF cohort (n=900) stratified by R-ISS (n=658) and the proposed CRSS (n=800) respectively. Median PFS for R-ISS1, R-ISS2 and R-ISS3 are 186, 151 and 79 weeks respectively with a p-value of 1.73e-5. Median PFS for CRSS-1, CRSS-2 and CRSS-3 are 249, 158 and 90 weeks respectively with a p-value of 8.64e-12. (C, D) Overall Survival (OS) in patients with MM from MMRF cohort (n=900) stratified by R-ISS (n=658) and the proposed CRSS (n=800) respectively. Median OS for R-ISS1, R-ISS2 and R-ISS3 are 264, Not reached and 164 weeks respectively. with a p-value of 6.58e-8. Median OS for CRSS-1, CRSS-2 and CRSS-3 are Not reached, Not reached and 238 weeks respectively with a p-value of 1.08e-15. (E, F) Univariate Cox hazard analysis on the prognostic factors. Hazard ratios for all the parameters except HRCA were calculated on complete data (n=900) for MMRF dataset. Hazard ratio for HRCA and the risk staging models were found using the data for which HRCA information was present (n=800 for MMRF dataset).

Table 5.7: Prediction of progression-free survival and overall survival (in %) for CRSS and R-ISS at 1, 2, 3, 4, and 5 years in the MMIn (n = 384) and MMRF datasets (n = 800).

	MMIn data						
	Year	**R-ISS (n = 355)**			**CRSS (n = 384)**		
		1	2	3	1	2	3
PFS	1	0.9318	0.8305	0.6967	0.8966	0.7812	0.7196
	2	0.8606	0.6601	0.5223	0.7709	0.6265	0.4472
	3	0.6404	0.5124	0.3632	0.6449	0.4729	0.2515
	4	0.3422	0.4179	0.2810	0.5251	0.3624	0.0587
	5	0.2738	0.2856	0.2342	0.4014	0.2679	0.0587
OS	1	0.9773	0.9387	0.7784	0.9630	0.8938	0.7976
	2	0.9540	0.8415	0.6393	0.9466	0.7679	0.6155
	3	0.9282	0.7764	0.5342	0.9098	0.6702	0.5831
	4	0.8895	0.6790	0.4953	0.8979	0.5691	0.4574
	5	0.8895	0.6422	0.3698	0.8979	0.4791	0.3136
	MMRF data						
	Year	**R-ISS (n = 658)**			**CRSS (n = 800)**		
		1	2	3	1	2	3
PFS	1	0.9033	0.8132	0.6358	0.9325	0.8367	0.6611
	2	0.7957	0.6261	0.4040	0.8162	0.6734	0.4423
	3	0.6295	0.4862	0.3059	0.7008	0.5084	0.3129
	4	0.4641	0.3414	0.2781	0.5151	0.3711	0.2249
	5	0.2769	0.2450	0.2781	0.4121	0.2637	0.1799
OS	1	0.9807	0.9092	0.8559	0.9869	0.9379	0.8231
	2	0.9612	0.8372	0.6460	0.9689	0.8772	0.6780
	3	0.9286	0.7799	0.5211	0.9478	0.8217	0.5814
	4	0.8833	0.7461	0.4904	0.9478	0.7844	0.5293
	5	0.5748	0.7108	0.3678	0.9478	0.6569	0.4691

of albumin, eGFR, and hemoglobin as the risk of disease increased (Figures 5.6 and 5.7) for both the MMIn and MMRF dataset. Wilcoxon-rank sum test was performed to compare the median values of the parameters between two successive risk groups and showed significant variation of parameters for both the datasets.

5.3.5 Model Interpretation

To ascertain the impact of individual parameters on risk stage predictions by CRSS, decision tree models built using MMIn and MMRF datasets were analysed using SHAP (Shapley Additive Explanations) (Figures 5.8 and 5.12). Key contributors of high risk predictions in the MMIn dataset were presence of HRCA, elevated levels of β2M, higher age and lower levels of albumin (Figure 5.8). Further, lower levels of eGFR and hemoglobin along with elevated levels of calcium also contributed to high risk pre-

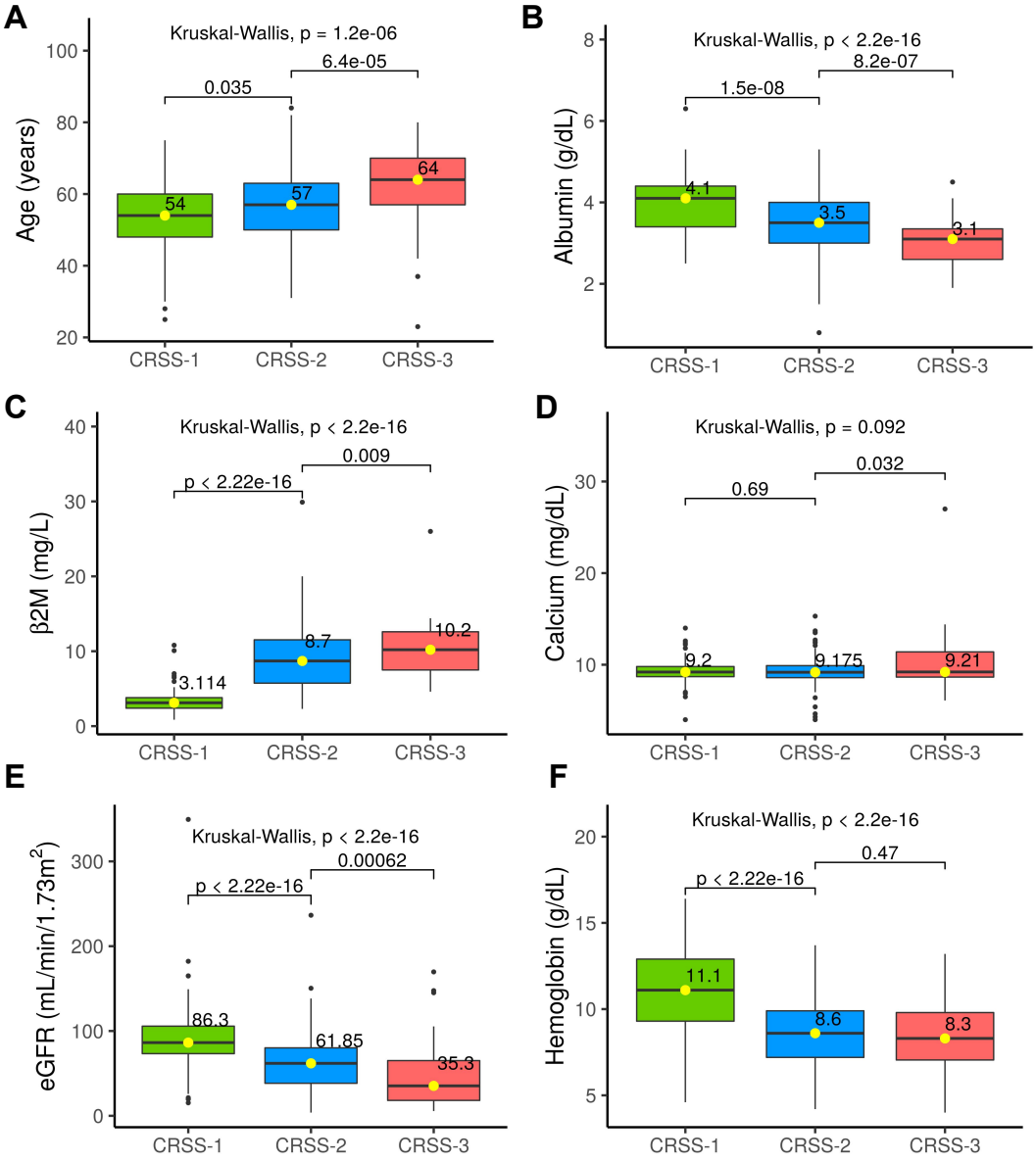


Figure 5.6: Boxplot showing the variation of the six parameters: A-age, B-albumin, C- β2M, D- calcium, E- eGFR and F-hemoglobin for MMIn dataset at CRSS-1, CRSS-2 and CRSS-3. The median values of all the parameters differ significantly across the three risk stages. Age and β2M are increasing while albumin, eGFR and hemoglobin are decreasing as the risk increases. Wilcoxon rank-sum test was used to compare two risk groups and Kruskal-Wallis test was used for comparing the three risk groups.

diction in the patients. It was observed from the waterfall plots (Figures 5.9, 5.10 and 5.11) of the randomly chosen patients in different risk stages that the order of the impact of the parameters varied in different patients within the same risk category. For the high-risk category (Figure 5.11), HRCA had the highest impact on one of the randomly chosen patients; in another patient, β2M had the highest impact in contributing to high

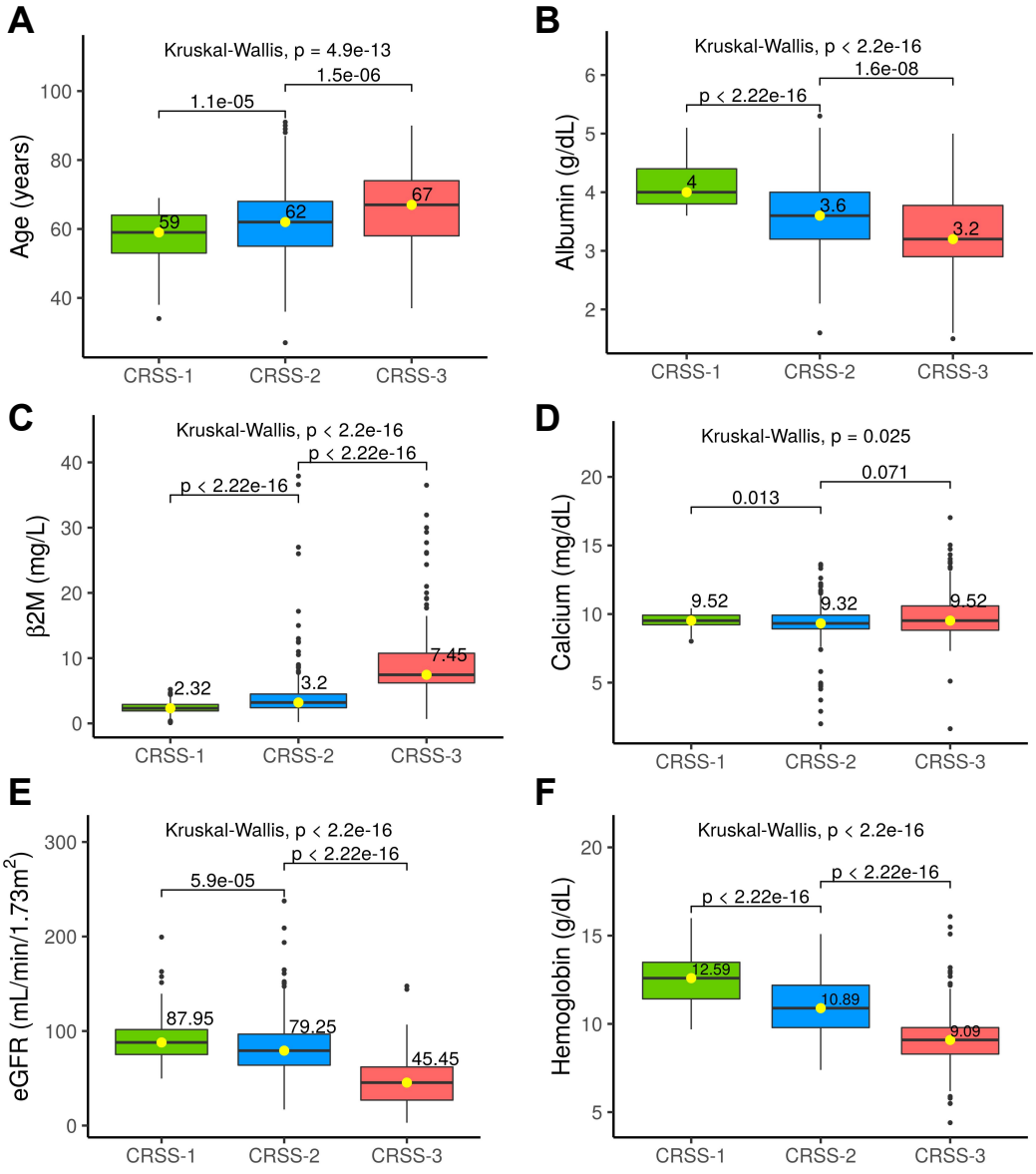


Figure 5.7: Boxplot showing the variation of the six parameters: A-age, B-albumin, C- β2M, D- calcium, E- eGFR and F-hemoglobin for MMRF dataset at CRSS-1, CRSS-2 and CRSS-3. The median values of all the parameters differ significantly across the three risk stages. Age and β2M are increasing while albumin, eGFR and hemoglobin are decreasing as the risk increases. Wilcoxon rank-sum test was used to compare two risk groups and Kruskal-Wallis test was used for comparing the three risk groups.

risk while in the third patient, age and albumin had the highest prognostic impact. This suggests that the risk assessment in MM is a cumulative function of multiple factors. An individual parameter cannot adequately capture the risk associated with MM given that other prognostic parameters could influence the outcome. Further, the complex association among different parameters that encapsulates the disease risk varies according to the patients, thereby, leading to varying order of impact of parameters in the

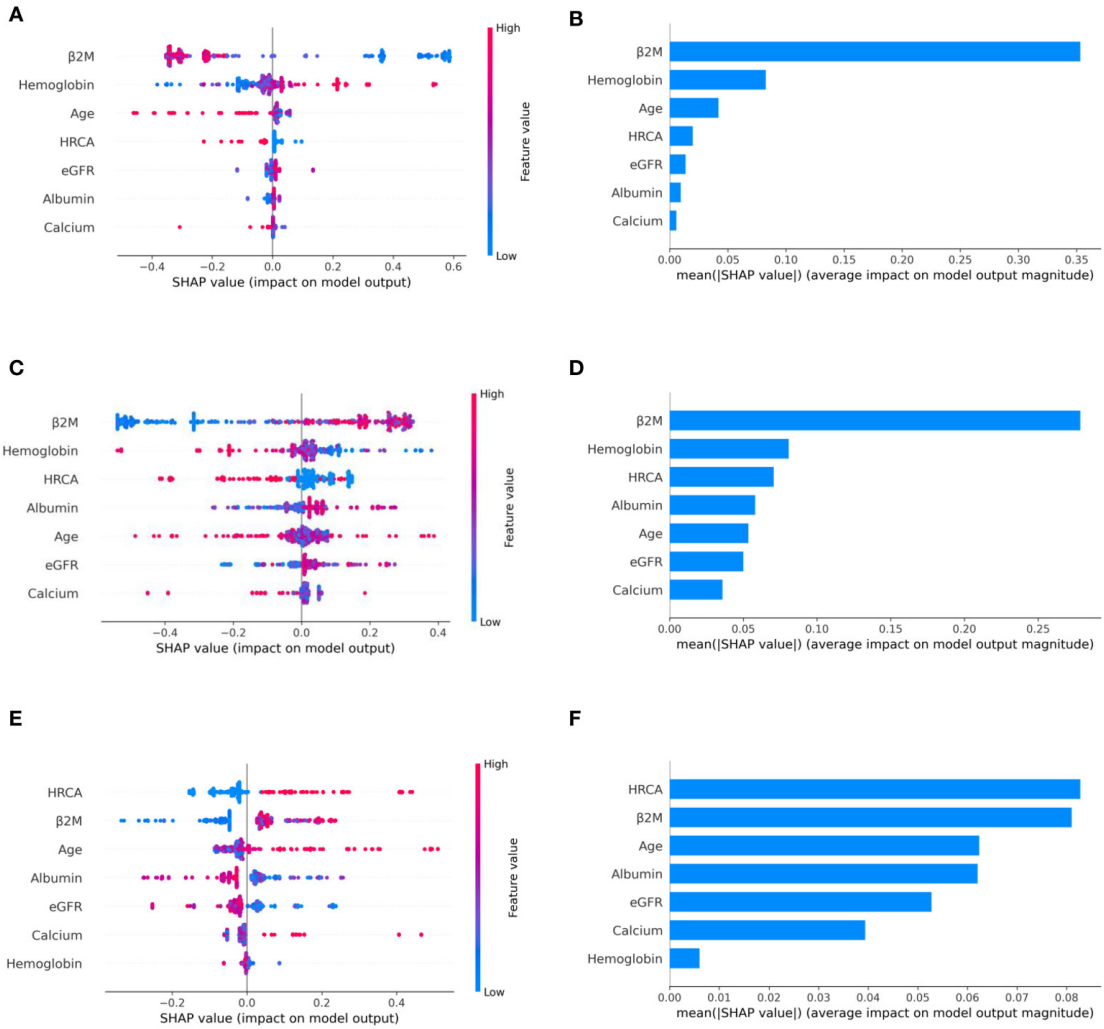


Figure 5.8: Model interpretation using SHAP (SHapley Additive exPlanations). SHAP summary plots for different risk stages inferred from MMIn data showing the relative impact of different parameters (top to bottom) contributing to a particular risk stage prediction. (A, B) CRSS-1: Normal levels of β2M and hemoglobin are the key contributors to the low-risk stage prediction. Furthermore, high values of age on the left side of the summary plot are pushing the model away from the low-risk prediction and are indicative of either intermediate or high risk. Overall, β2M has the highest impact and calcium has the lowest impact on the low-risk stage prediction. (C, D) CRSS-2: β2M and hemoglobin are the key contributors to the intermediate-risk stage. Elevated levels of β2M with lower levels of hemoglobin are indicative of intermediate risk. (E, F) CRSS-3: Presence of HRCA is contributing the most to the high-risk stage. Elevated values of β2M and calcium and lower levels of albumin, hemoglobin, and eGFR are contributing toward the high-risk stage prediction.

patients. Hence, the AI-based decision tree algorithms can handle such an integrated analysis. This analysis reveals that each patient is unique and multiple factors interact and impact the outcome differently in individual patients.

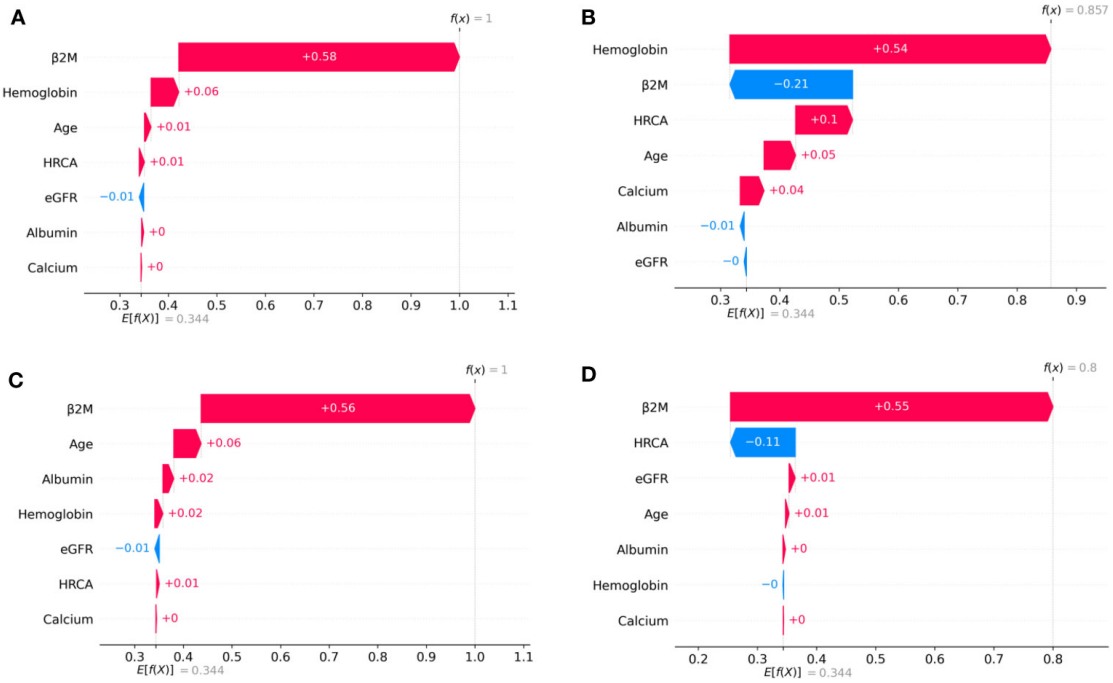


Figure 5.9: SHAP waterfall plots for the randomly chosen four patients in low-risk stage (CRSS-1) from the MMIn dataset. The pink color shows the positive impact of the feature, while the blue color shows the negative impact of the feature. Features with a positive impact contributed to the class of low-risk stage prediction, while features with a negative impact contributed to class opposite to low risk. β2M, hemoglobin, age, and HRCA have the highest overall impact on low-risk stage prediction in the MMIn dataset. However, this ranking itself differs from patient to patient as can be seen in (A–D). (A) β2M has the highest impact followed by hemoglobin, age, and HRCA. (B) Hemoglobin has the highest impact followed by β2M and age. (C, D) β2M has the highest impact followed by age and HRCA.

5.4 Discussion

The influence of ethnicities on clinical characteristics in patients belonging to distinct ethnic groups is well known and therefore, it is of paramount interest to integrate the ethnic group specific information in risk-staging models as it can affect the risk score prediction. R-ISS3 is the current standard of care for staging myeloma patients which includes a few HRCA but molecular aberrations such as 1q gain and chromothripsis associated with adverse outcome have been overlooked [197]. In fact, it includes t(4;14) which has lost significance in patients treated with triplet regimens [198]. Besides, R-ISS does not include any ethnic specific information and therefore, is not robust considering the large heterogeneous population of MM patients globally. An ideal risk staging system would be based on all the known adverse prognostic factors including clinical, ethnic and molecular aberrations. There is tremendous heterogeneity in global healthcare systems that limit availability of high end molecular testing for all patients and yet, the internet/electronic connectivity allow patients to receive medical

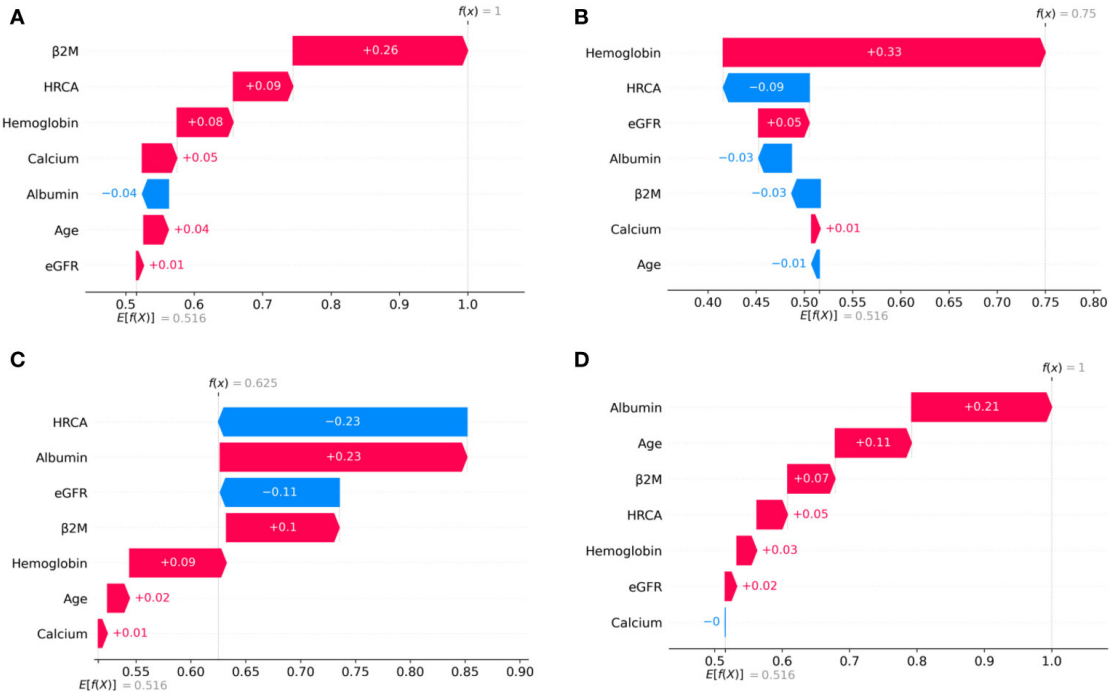


Figure 5.10: SHAP waterfall plots for the randomly chosen four patients in the intermediate-risk stage (CRSS-2) from the MMIn dataset. The pink color shows the positive impact of the feature, while the blue color shows the negative impact of the feature. Features with a positive impact contributed to the class of intermediate-risk stage prediction, while features with a negative impact contributed to the class opposite to intermediate risk. β2M, hemoglobin, HRCA, and albumin have the highest overall impact on the intermediate-risk stage prediction in the MMIn dataset. However, the ranking of the features itself differs from patient to patient as can be seen in (A–D). (A) β2M has the highest impact followed by HRCA. (B) Hemoglobin has the highest impact followed by HRCA. (C) HRCA has the highest impact followed by albumin. (D) Albumin has the highest impact followed by age.

advice from global leaders in medicine. Recently, an AI-supported risk staging model, MRS [199], has been developed for NDMM, however, it does not include HRCA and ethnicity information. Considering the present world scenario, it is, thus, desirable to develop a simple risk-staging model that integrates ethnic specific characteristics of the prognostic parameters that are easy to acquire in the healthcare settings worldwide.

5.4.1 Risk-staging models and their performance as compared to R-ISS

In contrast to R-ISS which utilizes four parameters, seven parameters were taken into consideration for designing CRSS. It was observed that the cut-off values for these parameters derived using KAP, vary in the two cohorts, one of which belongs to Indian and the other belongs to the American population. For the Indian data, there was a

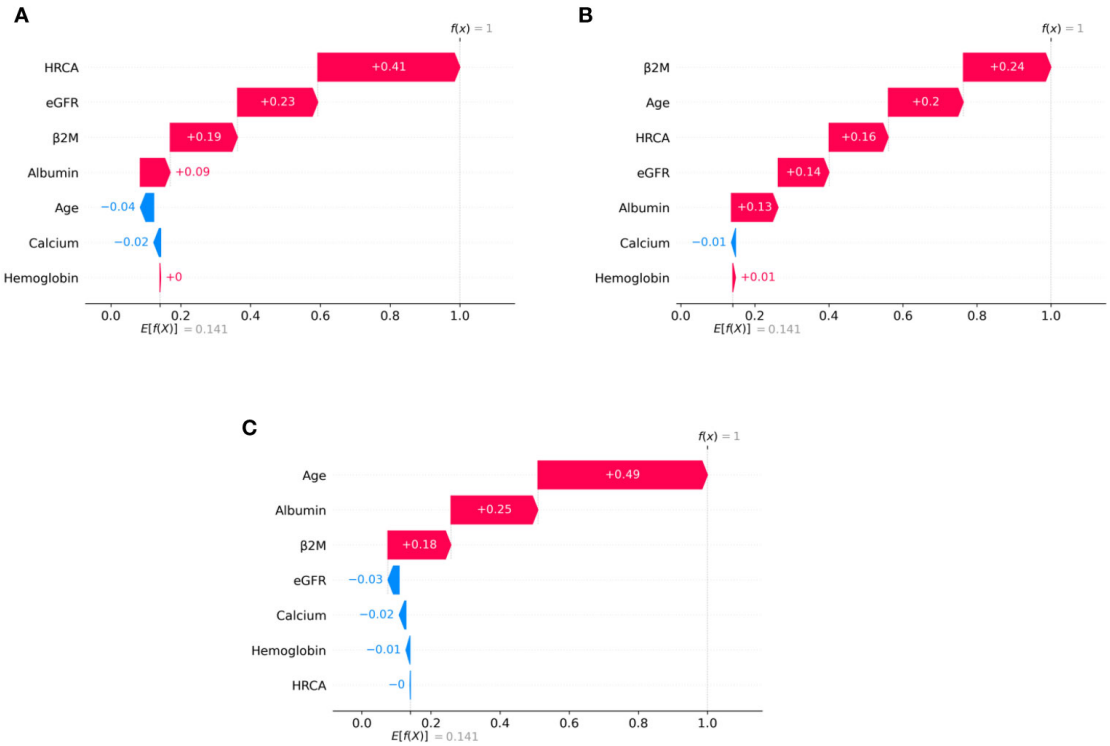


Figure 5.11: SHAP waterfall plots for randomly chosen patients in high-risk stage (CRSS-3) from the MMIn dataset. The pink color shows the positive impact of the feature, while the blue color shows the negative impact of the feature. Features with a positive impact contributed to the class of high-risk stage prediction, while features with a negative impact contributed to class opposite to highest risk. HRCA, β2M, age, and albumin have the highest overall impact on high-risk stage prediction. However, this ranking differs from patient to patient as can be seen in (A–C). (A) HRCA has the highest impact. (B) β2M has the highest impact. (C, D) Age and albumin have the highest impact.

change in the cut-off values for β2M, age, eGFR and hemoglobin while there was no change in the cut-off value for calcium and albumin as shown in (Table 5.2). For MMRF data, there was change in cut-off values for calcium, eGFR, hemoglobin and age while the cut-off values for albumin and β2M remain unchanged. The median age of onset of MM in the Indian population is almost a decade early as compared to the population in the USA [200, 201]. This supported our assertion of choosing different cut-offs of age for MMIn from the MMRF dataset.

Various models were built on the different combinations of the parameters using both the established and proposed cut-offs for the two datasets. The best staging model for both the dataset was obtained when proposed cut-offs for the respective cohorts were used. When the ML-derived cut-offs were used for the parameters age, eGFR, hemoglobin and β2M in A3 model, performance was enhanced significantly in terms of high C-index and hazard ratios as compared to R-ISS. A similar observation was noticed in the M4 model which utilized ML derived cut-offs obtained for MMRF dataset and

achieved the best performance among all the models with a significant improvement in the C-index as well as hazard ratios as compared to R-ISS. Overall, A3 and M4 were the best staging models for MMIn and MMRF data respectively. The improvement in the performance of the model verified our hypothesis that the cut-offs of the different parameters vary with different ethnicities.

The plausibility of the proposed model was further substantiated by performing significance testing. Kruskal-Wallis test showed statistically significant variations ($p < 0.05$) in the median values of the parameters-age, albumin, β2M, eGFR, hemoglobin across the three risk groups (Figures 5.6 and 5.7) for both the datasets. Further, Wilcoxon rank-sum test revealed statistically significant variations ($p < 0{\cdot}05$) in the median values of the parameters between two successive risk groups (CRSS-1 and CRSS-2; CRSS-2 and CRSS-3). Further, CRSS for MMIn and MMRF dataset were interpreted using SHAP (Shapley Additive Explanations) to establish the clinical relevance of the risk stages predicted by CRSS. For MMIn data, elevated levels of β2M and calcium with lower levels of eGFR and hemoglobin contributed to high risk whereas in MMRF data, elevated levels of β2M and lower levels of hemoglobin, eGFR and albumin contributed to high risk in myeloma patients. These findings are in accordance with the observations mostly identified in high risk MM patients. Additionally, it was observed that the order of impact of hemoglobin was higher in low risk stage prediction in MMIn dataset as compared to MMRF dataset while the order of impact of hemoglobin was higher in high risk stage prediction in MMRF dataset as compared to MMIn dataset (Figures 5.8 and 5.12). The difference in the rankings can be attributed to the varying ethnicities and further confirmed our claim of using ethnicity-aware risk staging models for MM. In the present study, we have used MMIn and MMRF cohorts belonging to Indian and American ethnicities respectively for building CRSS models. Results on both the cohorts have strengthened our claim that the robustness of the staging model is amplified by inclusion of ethnicity-specific cut-offs of the prognostic factors as well as by utilizing AI techniques.

The classification rules were obtained using a Decision tree classifier on the classification output of the best performing models in both MMIn and MMRF data. Overall classification accuracy was 94.79% and 98% for the MMIn and MMRF data respectively. Final risk-stages were evaluated using the classification rules in both the dataset. Further, it is evident from the UMAP plots that both the MMIn and MMRF data were not visible as three separate risk groups initially in the absence of CRSS risk labels (Figures 5.13A, 5.13C and 5.13E). With the addition of these risk labels with every patient sample, the subjects could be seen to be grouped separately (where a group corresponds to one risk label) in the UMAP plot (Figures 5.13B and 5.13D). This demonstrates the ability of the CRSS model in identifying the risk groups correctly from the non-separable data. To further validate our model, we found risk stages in 123 prospective subjects of

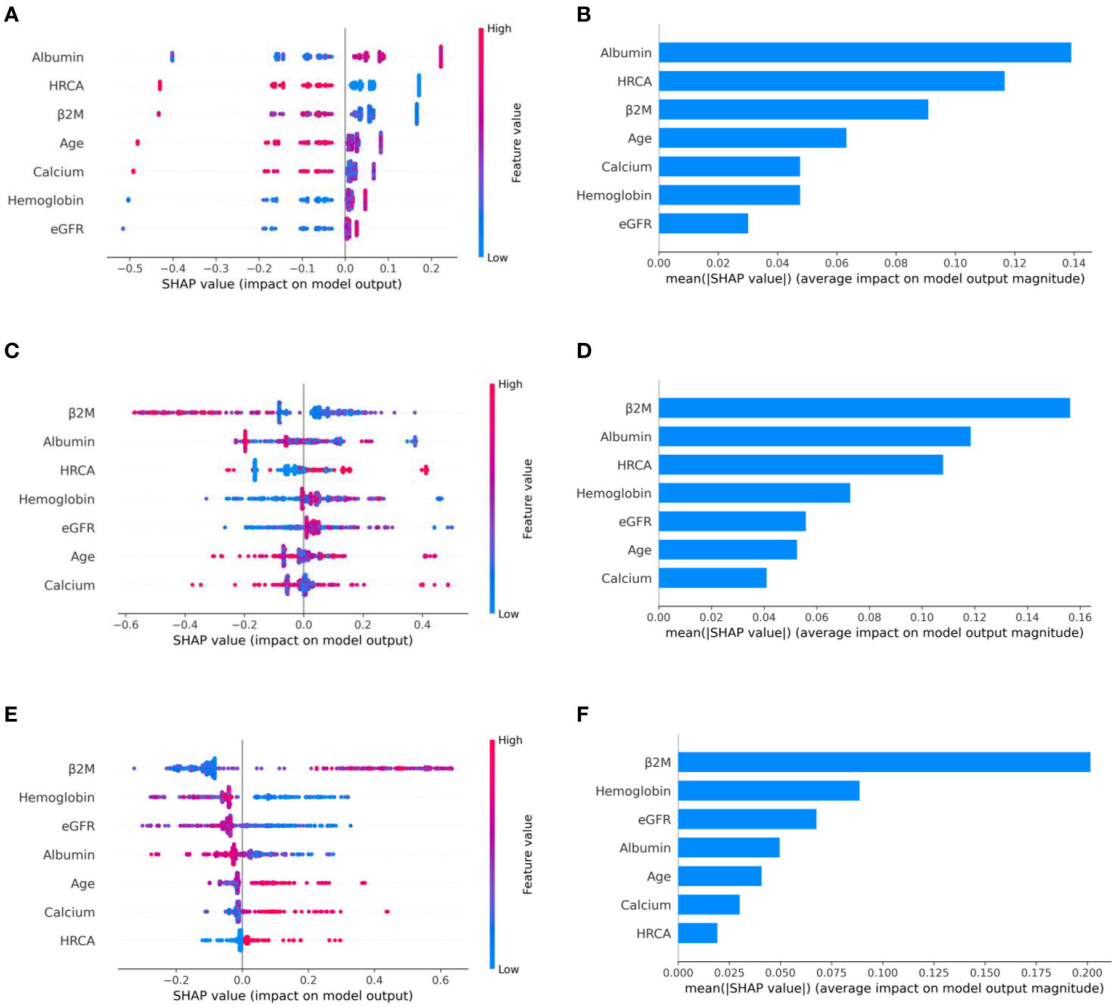


Figure 5.12: Model interpretation using SHAP. SHAP summary plots for different risk stages inferred in MMRF data showing the impact of different parameters used in the model. (A, B) CRSS-1: albumin, HRCA, and β2M have the highest impact on the low-risk stage. Normal levels of albumin, absence of HRCA, and lower values of β2M are contributing to low risk (CRSS-1) in myeloma patients. (C, D) CRSS-2: β2M, albumin, and HRCA are the key contributors to the intermediate-risk stage. (E, F) CRSS-3: β2M and hemoglobin have the highest impact on the high-risk stage. Elevated levels of β2M and lower values of hemoglobin are contributing toward the high-risk stage in the patient. Lower values of albumin and eGFR are further promoting high-risk stage prediction.

MMIn data that were not used to build the CRSS model. UMAP plots (Figure 5.13F) suggest that the prospective subjects got correctly aligned to their respective risk stages inferred via CRSS.

For MMIn data, β2M was in the highest level of hierarchy in the classification rules followed by hemoglobin and HRCA (Figure 5.3(a)). For MMRF data, the prognostic factor in the highest level of hierarchy was β2M followed by albumin and Hb (Figure 5.3(b)). The cut-off values for β2M, albumin and Hb were 5.2, 3.55 and 9.64. The cut-

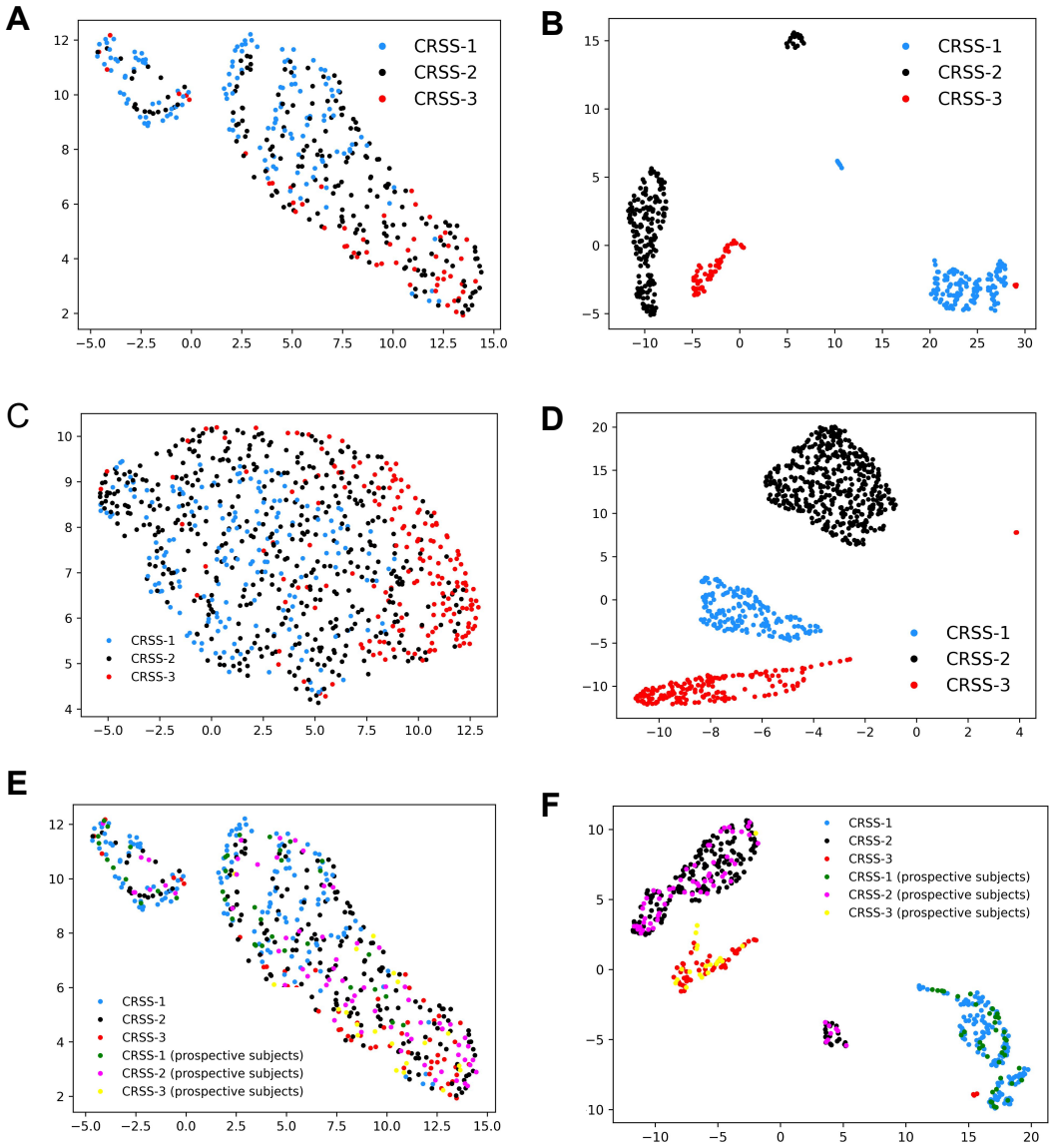


Figure 5.13: UMAP scatter plot of (A), (B) MMIn data and (C), (D) MMRF data depicting the data in absence and presence of risk stage labels respectively. The plot indicates that both the MMIn and MMRF data were not visible as three separate risk groups initially in the absence of CRSS risk labels. With the addition of these risk labels, the patients are now grouped separately (where a group corresponds to one risk label) in the UMAP plot. This demonstrates the ability of the CRSS model in identifying the risk groups correctly from the non-separable data. Performance of the model was further validated by identifying risk stages in 123 prospective MMIn subjects that were not used to build CRSS. (E) UMAP scatter plot of the prospective MMIn subjects (n=123) along with the MMIn data of 384 patients reveals that data is not visible as separate risk groups in absence of risk stage labels and (F) UMAP scatter plot reveals that the prospective MMIn subjects align themselves to their respective risk groups after addition of risk stage labels.

offs for β2M and albumin were not changed but the cut-off value proposed for Hb was 9.59 which was close to the observed value in the classification rules. This observation further justified our choice of using new cut-offs for the risk-staging model.

5.4.2 Conclusion

In this work, we examined the impact of ethnicity based cut-offs of laboratory parameters derived using ML algorithm on risk prediction in Indian and American patients with MM. We trained different risk staging models for both MMRF and MMIn dataset. The best predictor model was obtained when ethnicity specific cut-offs of the clinical parameters were utilized. Further, we presented a new reliable and robust AI-enabled risk staging system, namely, CRSS that utilizes easily acquirable laboratory and clinical parameters i.e. age, albumin, β2-microglobulin (β2M), calcium, estimated glomerular filtration rate (eGFR) and hemoglobin along with HRCA. Risk-stratification achieved by AI-assisted CRSS is able to better separate the patients into different risk groups as compared to R-ISS. High concordance-index and hazard ratios reveal the superior performance of CRSS as compared to R-ISS. Further, the clinical and biological significance of the decision tree classifier rules for risk stage prediction in MM patients was deduced via SHAP analysis on both the datasets. The successful evaluation of our proposed staging system on both the datasets establishes the utility of the proposed ethnicity aware staging system for NDMM patients, treated largely with novel agents or a combination thereof, in a real-world scenario. Our study also highlights the importance of application of AI in building CRSS thereby enhancing the prediction of survival outcome and separability of risk-stages in NDMM patients. We have also developed a web platform based AI-assisted ethnicity aware MM risk staging calculator. Screenshot of the online calculator is shown in the Figure 5.14.

5.4.3 Limitations and Future work

CRSS has been built on a smaller set of NDMM patients as compared to the R-ISS3 study. In future, the CRSS model may be tested on larger datasets with varying ethnic groups as the cohort size of the present study is 25% of the cohort used in R-ISS reported in 2015. As the CRSS calculator becomes available online, data could be generated by independent groups for further validation in real world scenarios.

Consensus based Risk Staging System (CRSS) calculator for Multiple Myeloma (version 1.0)

Joint collaborative work of Laboratory Oncology Unit, Dr. B.R.A. IRCH, AIIMS, New Delhi and SBILab, Department of ECE, IIIT-Delhi, New Delhi

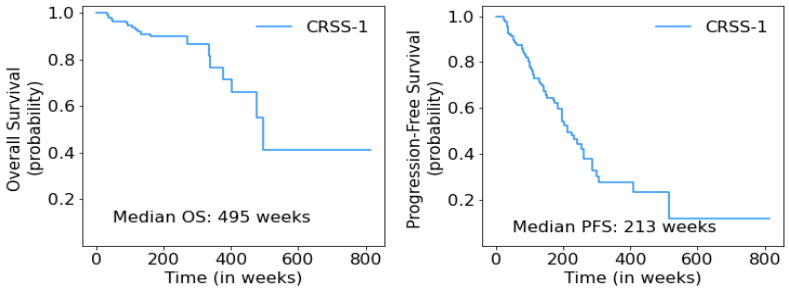
Principal Investigators: Prof. Ritu Gupta, AIIMS and Prof. Anubha Gupta, IIITD

Description: An efficient and robust AI-enabled risk-staging system for MM patients that utilizes ethnicity-specific cutoffs of key prognostic parameters. It predicts the risk stage of a patient depending on the values of the seven parameters- age, albumin, β2m, hemoglobin, calcium, eGFR and high risk cytogenetic abnormalities [del 17p; t(4;14); t(14;16)].

- It's utility has been validated for Newly diagnosed Multiple Myeloma (NDMM) patients.
- Risk-stratification achieved by AI assisted CRSS is able to better separate the patients into different risk groups as compared to RISS
- It is a reliable and efficient tool for upfront risk stratification of MM patients and can help the clinicians/doctors in designing and providing effective therapy to MM patients.

Ethnicity(Indian/American)	◉ Indian	○ American
Age (years)	45	
Albumin (g/dL)	4	
β2M (mg/L)	4	
Calcium (mg/dL)	10	
eGFR (mL/min)	78	
Hemoglobin (g/dL)	12	
High risk cytogenetic abnormalities	◉ No	○ Yes
	Reset Calculate	
CRSS stage	1	

Low Risk



Please cite us if you use CRSS calculator in your research work.

- Farswan A, Gupta A, Sriram K, Sharma A, Kumar L, Gupta R. Does ethnicity matter in multiple myeloma risk prediction in the era of genomics and novel agents? Evidence from real-world data. Front Oncol 2021.

If information on ethnicity and genetic abnormalties is not available, then you can use MRS calculator designed by us. MRS is also an advanced AI-supported calculator that works efficiently in the absence of cytogenetic abnormalities, that is, it predicts the MM cancer risk stage using the six parameters of the patients- age, albumin, hemoglobin, β2M, calcium and eGFR.

Figure 5.14: Online version of CRSS calculator

Chapter 6

Inference of clonal trajectory in single-cell data

6.1 Introduction

Clonal heterogeneity is an established feature of cancer, characterized by the co-existence of genetically divergent clonal sub-populations of malignant cells within the tumor [202, 203]. The cause of the intra-tumor heterogeneity is genomic instability [204] which promotes elevated mutation rate [205] by the sequential acquisition of somatic mutations within the diverse subclones. The multiple subclones differ in their immunological characteristics, growth rate, and ability to metastasize. Further, the treatment therapy affects every clone distinctly due to the diversified nature of the subclones. A group of the clones extinct, while others become resistant to the drug/therapy, eventually causing relapse in the cancer patients. Multiple studies have established the fact that intra-tumor heterogeneity adversely affects the overall drug response in cancer patients [206]. It is, therefore, critical to gain insight into the process of cancer evolution and characterize the intra-tumor heterogeneity as it may foster the development of drugs to deal with the treatment resistance clones in the cancer patients [205]. Next-generation sequencing technology has enabled the identification of the genomic changes in the tumor population at a higher resolution. Several computational methods have been proposed to infer the pattern of clonal evolution from the high throughput next-generation bulk sequencing data [207, 142, 208, 209, 141]. However, reconstructing the clonal evolution pattern from bulk DNA sequencing data is difficult because the bulk DNA data comprises of the mixtures of mutations from thousands to millions of heterogeneous cells in the sample.

With the advent of single-cell sequencing (SCS) technology, the resolution of tumor cell profiling has vastly improved, thereby leading to a better quantification of intratumoral heterogeneity [210, 211]. Several methods for constructing tumor phylogenies have been developed that utilize single-cell copy number variation (scCNV) profiles or single-cell single nucleotide variation (scSNV) data. For deriving cancer evolution from scSNV data, distance-based methods like UPGMA were initially proposed in [212, 213]. In [214, 215], copy number profiles from scCNV data were employed to reconstruct tumor phylogeny. Single-cell sequencing data helps overcome the struggle of inferring tumor phylogeny from the bulk sequencing data; it poses other challenges arising during the sequencing process. These challenges comprise of several errors found in the single-cell data. The most common types of errors are false positives (FP),

false negatives (FN), missing bases (MB), and doublets. False positives arise when a mutation is called, but it is not present in the cell. A false negative is created when either one or both the alleles do not amplify enough during allele dropout events and are thus neglected in the final computation. Doublets occur when two or more cells are measured simultaneously due to errors in cell capture or sorting. It can be deduced from multiple studies that the false positive rate is in the range of 10^{-5} while false negative varies from 0.1 to 0.43 [212, 216]. Cell doublets have varying rates of occurrence depending on the technique used to isolate cells and can range from 1% to 10% [217]. Further, single-cell data also suffers from the problem of missing bases which arise out due to insufficient coverage of the signal during the sequencing process. The highest reported missing rate is 58% [215, 218]. Multiple methods have been proposed for the imputation of the missing values in the genomics data, such as DSNN [190].

Errors associated with single-cell data can adversely affect the downstream analysis. Therefore, numerous methods have been devised to resolve these errors in single-cell data in recent times. Examples include SCITE by [219], OncoNEM by [220], SCG by [221] and SciΦ by [222]. However, most of these methods are based on the assumptions of the infinite site model (ISM); though few of the methods do account for the loss of heterozygosity (LOH), they completely overlook the recurrent mutations. In the ISM model, a locus in the cell has the presence of mutation, denoted by 1 or the absence of mutation, marked by 0. Transitions between these states are restricted so that a mutation can be gained atmost once during the cancer evolution, and it cannot be lost after it is gained. The phylogeny constructed by following the principles of the infinite-sites model is known as a perfect phylogeny. When the data is error-free, a unique and perfect phylogeny is created. However, if the data has errors, it may lead to multiple phylogenies as there are numerous ways of removing the errors from the single-cell data. Further, the assumption of the infinite site model may not always hold for tumor evolution where simultaneously multiple recurrent events occur ([223]. Therefore, methods like SiFit [217], BEAM [224] and RobustClone [225] have been developed that place no constraints on the mutation model. All the methods perform well for the dataset of small to medium size, but they become computationally expensive for large datasets. Methods like OncoNEM fail in computation on the dataset of median size 500×500 [225]. Further, the size of the single-cell data is increasing with rapid evolution in single-cell technology, ultimately leading to enhanced computational complexity. Therefore, a robust and computationally inexpensive method that works well on large and small datasets needs to be formulated.

Recently, Robust Principal Component Analysis (RPCA) was utilized to recover denoised matrix from the observed noisy data [225]. RPCA decomposes the observed noisy matrix into a low-rank matrix and a sparse error matrix, the noise associated with the observed matrix. Clones were inferred from the denoised matrix via the Louvain-

Jaccard algorithm, and ultimately, a clonal tree was deduced from the clones via a minimum spanning tree algorithm. However, the Louvain method may yield poorly connected clusters in the data, as established by [226], therefore the authors proposed Leiden algorithm that outperforms Louvain by determining better partitions in the data. RobustClone uses the Louvain-Jaccard algorithm, and hence, it may not always infer the optimal number of clusters. Further, the method infers subclones via the denoised matrix, and if the matrix is not denoised properly, it may again lead to an inaccurate number of clusters in the data. Therefore, we propose a novel method, ARCANE-ROG (Algorithm for Reconstruction of CANcer Evolution via RObust Graph Learning), to infer evolutionary cancer patterns from single-cell DNA data. The first step involves a robust graph learning-based method [227] which denoises and imputes the noisy and incomplete matrix. The original algorithm [227] worked only on noisy data and was evaluated on image data. In our work, we have improved the algorithm to impute the missing entries in addition to denoising. An adjacency graph is also learnt from the data simultaneously. Both the denoising and adjacency graph learning operations boost each other such that the overall performance of the denoising algorithm is improved. In the second step, the learned graph is used to infer an optimal number of clusters/subclones in the data via the Leiden algorithm. The adjacency graph is used instead of the denoised data for inferring subclones which ensures error-free prediction of subclones. Finally, a clonal pattern is deduced using subclonal information via a minimum spanning algorithm. We have compared the performance of ARCANE-ROG with RobustClone [225], BnpC [228] and GRMT [229]. The novel contributions of the work are as follows.

- Robust graph learning is utilized to denoise and impute the noisy and incomplete binary matrix of single-cell DNA data. An adjacency graph is simultaneously learnt from the input matrix.
- Learned adjacency matrix is used to infer the number of subclones in the data instead of the denoised matrix.
- Optimal number of subclones are identified in the data via the Leiden algorithm.

The work has been organized in the following manner. The methods section describes the method and the algorithm proposed in the current study, along with a detailed description of the simulated and real datasets used to validate our proposed method's performance. The results and discussion section give an exhaustive illustration of the performance of our proposed ARCANE-ROG method on the simulated and real datasets. It also shows the comparison of ARCANE-ROG with the state-of-the-art RobustClone method in terms of different evaluation metrics. Finally, we conclude our proposed method by establishing the robustness and the efficiency of the proposed method.

6.2 Methods

Our proposed method, ARCANE-ROG, is based on simultaneously learning the adjacency graph from the data while denoising and imputing the noisy and incomplete data. Leiden algorithm is then applied on the learned graph to find the optimal number of clusters, and finally, the clonal tree is inferred via a minimum spanning tree algorithm. In this section, we first explain how the algorithm was extended for accommodating missing entries and then, we describe the Leiden algorithm and Minimum Spanning Tree algorithm used for recovering the number of clones and clonal tree, respectively. Robust Graph learning [227] method to denoise the noisy binary input data matrix has been explained in the Appendix C.

6.2.1 Proposed Extension of Robust Graph Learning for Recovering Missing Values

In this work, we have extended robust graph learning for handling missing entries in the data. Consider a noisy single-cell data matrix, X^N, of size $m \times n$. m is the number of cells and n is the number of mutation sites. This matrix is binary in nature where '1' represents the presence of mutation and '0' denotes the absence of mutation. Our goal is to extract a denoised matrix, X^D from this noisy matrix along with E which is the error observed in matrix such that $X^N = X^D + E$. This task of denoising a noisy matrix is very well performed by applying robust PCA on the noisy matrix as done in [225]. Low rank constraints are added on the denoised matrix, X^D, because the original genotype matrix is a low rank matrix where the tumor cells are clustered together into various subclones such that there is little to no variation in the genotype of the cells within the same subclone. Further, we consider that error component, E, is sparse. Single-cell data not only suffers from noisy corruptions but it also has missing values which needs to be tackled for accurate downstream analysis. Therefore, a linear operator, $P_\Omega(X^D)$, was defined which sets the unobserved entries to 0 while keeping the rest equal to the observed entries as follows:

$$P_\Omega(X^D) = \begin{cases} X^D_{ij} & \text{if(i,j)} \in \Omega; \\ 0 & \text{if(i,j)} \notin \Omega. \end{cases} \tag{6.1}$$

Our objective now is to recover a denoised matrix, recover missing values as well as to learn an adjacency graph during the denoising process. Consider L and S to be the Laplacian and similarity graph learned during the denoising step. The objective is

formulated in equation 6.2.

$$\begin{aligned} \min_{X^D,E,S} &||X^D||_* + a||E||_1 + b\text{Tr}X^D L(X^D)^T + c||S||_F^2, \\ \text{s.t.} \quad & P_\Omega(X^N) = P_\Omega(X^D + E), S1 = 1, 0 \leq S \leq 1 \end{aligned} \tag{6.2}$$

where a, b and c are the trade-off parameters. In the above formulation, denoising, and graph learning are implemented together such that each of the step iteratively enhances the other step. The above formulation ensures that the low rank denoised matrix and noisy sparse component is recovered from the observed entries, $P_\Omega(X^N)$. The above equation can be further converted to the following framework:

$$\begin{aligned} \min_{X^D,E,S} &||X^D||_* + a||P_\Omega(E)||_1 + b\text{Tr}X^D L(X^D)^T + c||S||_F^2, \\ \text{s.t.} \quad & X^N = X^D + E, S1 = 1, 0 \leq S \leq 1 \end{aligned} \tag{6.3}$$

W is an auxiliary variable and the above formulation can now be solved via alternating direction method of multipliers (ADMM) by adding an auxiliary variable W in the equation.

$$\begin{aligned} \min_{X^D,E,S,W} &||X^D||_* + a||P_\Omega(E)||_1 + b\text{Tr}(WLW^T) + c||S||_F^2, \\ \text{s.t.} \quad & X^N = X^D + E, S1 = 1, 0 \leq S \leq 1, W = X^D \end{aligned} \tag{6.4}$$

Augmented Lagrangian function can be obtained by removing equality constraints on X^N and W:

$$\begin{aligned} \mathcal{L}(X^D, E, S, W, Z_1, Z_2) = &||X^D||_* + a||P_\Omega(E)||_1 + b\text{Tr}(WLW^T) + c||S||_F^2 \\ &+ \frac{\mu}{2}\left(||X^D + E - X^N + \frac{Z_1}{\mu}||_F^2 + ||X^D - W + \frac{Z_2}{\mu}||_F^2\right) \\ &\text{s.t.} \quad S1 = 1, 0 \leq S \leq 1 \end{aligned} \tag{6.5}$$

where μ is penalty parameter and Z_1 and Z_2 are the Lagrangian multipliers. In our proposed method, ARCANE-ROG, a was set to $(1 + 3 \times \Omega)/\sqrt{m \times n}$, b was set to $5/\sqrt{m \times n}$ and c was set to $5/\sqrt{m \times n}$. The steps to solve the above problem and the algorithm used for denoising the data has been provided in the Appendix C.

6.2.2 Leiden algorithm

Leiden algorithm is a type of community/cluster detection method proposed in [226] for the analysis of complex networks. One of the best performing community detection algorithms in the literature is the Louvain algorithm [230]. However, there is a major problem associated with Louvain, i.e., it may yield arbitrarily poorly connected and sometimes internally disconnected communities/clusters. Hence, the Leiden algorithm [226] was proposed to resolve the shortcomings posed by the Louvain algorithm. It is based on an algorithm introduced in [231], which was an improvement of the Louvain algorithm. In addition, the Leiden algorithm has also utilized ideas proposed in [232, 233, 234] to improve its performance. It has been shown in the paper [226] that the Leiden algorithm significantly outperforms the Louvain algorithm. Leiden algorithm guarantees well-connected communities/clusters. Leiden algorithm is explained as follows (Figure 6.1). Consider an undirected graph, $G = (V, E)$ with nodes, $n = |V|$ and edges, $m = |E|$. A partition is defined as $\mathcal{P} = \{C_1, C_2, ...C_r\}$ where $r = |\mathcal{P}|$ denotes the number of communities. Each community C_i is a subset of V and consists of a set of nodes such that $V = \bigcup_i C_i$ and intersection of C_i and C_i is an empty set ($C_i \cap C_i = \emptyset$) for all when $i \neq j$. To ascertain the quality of the partition, a quality function, $\mathcal{H}(G, \mathcal{P})$ is defined as follows.

$$\mathcal{H}(G, \mathcal{P}) = \sum_{C \in \mathcal{P}} \left[E(C, D) - \eta \binom{||C||}{2} \right], \tag{6.6}$$

where η is the resolution parameter and $E(C, D)$ represents the number of edges formed between the nodes in the communities denoted by C and D. $||C||$ is the cardinality of the flattened set C i.e $||C|| = |flat(C)|$. Our objective is to find the highest possible quality partition. We start with an initial partition P_0 which is usually the singleton partition of the graph G, i.e each nodes act as an individual community ($P_0 = \{\{v\}|v \in V\}$). Now nodes are moved from one community to another to find a partition, $\mathcal{P}$. This partition is then refined to create a refined partition, $\mathcal{P}_{\text{refined}}$. The refinement phase is unique to the Leiden algorithm. During this phase, nodes are merged with the community randomly, and the community with the largest increase in the quality function, $\mathcal{H}(G, \mathcal{P})$, is selected. Thus, communities in $\mathcal{P}$ may split into multiple sub-communities in $\mathcal{P}_{\text{refined}}$. In the Figure 6.1, the magenta community in (B) is refined into two sub-communities shown by red and magenta in (C), which are separated into two nodes after aggregation in (D); however, they belong to the same community. Once an aggregate network is created from the refined partition, individual nodes are further moved in the aggregate network and refined. This procedure is repeated till no further improvements in the partitions can be made. More details on the method and its code can be found in the paper [226].

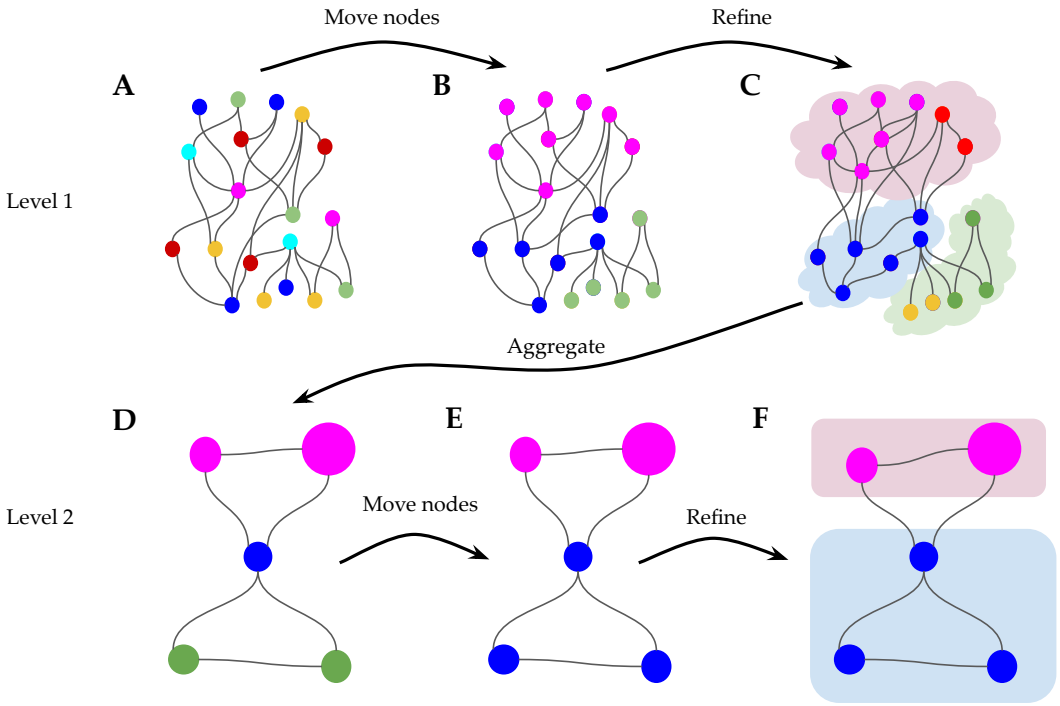


Figure 6.1: The Leiden algorithm starts with an initial partition which is usually the singleton partition of the graph, i.e. each nodes act as an individual community (A). Individual nodes are moved from one community to another to find an initial partition (B). The initial partition so formed is then refined as in (C). During the refinement phase, nodes are merged with the community randomly, and the community with the largest increase in the quality function is selected. Thus, communities in the initial partition may split into multiple sub-communities in the refined partitions. An aggregate network (D) is formed using the refined partitions. It is to be noted that the aggregate network is initially created using the non-refined partition. Individual nodes are then moved in the aggregate network (E). Refining the aggregate network may or may not change the partition. There is no change in the partition (F) in this case. These steps are repeated until there is no scope for further improvement.

6.2.3 Minimum Spanning Tree

Minimum spanning tree (MST) can be described as a spanning tree having the minimum cost among all spanning trees. For example, consider the graph G in 6.2(A). A graph can be represented as $G(V, E, w)$ where V represents vertices A, B, C, D, E, and E represents edges and w represents weights of the edges 1, 2, 3, 4, 5, 6, 7, and 8. A spanning tree of the graph can be built if it satisfies two main conditions, 1) the number of vertices in the tree is equal to the number of vertices in graph G, and 2) the number of edges present in the spanning tree is a subset of the number of edges in G. There can be multiple spanning trees corresponding to a graph as shown in Figure 6.2(B-E). We can compute the cost of each tree by the addition of weights of all edges in the spanning tree. The tree having the minimum cost is defined as the minimum spanning

tree. For graph G, MST is shown in the Figure 6.2(E) having the lowest of cost of 10 as compared to other trees having cost 12 (6.2(B)), 14 (6.2(C)), and 22 (6.2(D)). There can also exist multiple minimum spanning trees to a graph. Minimum spanning tree algorithm is extensively used in problems such as image segmentation, cluster analysis etc.

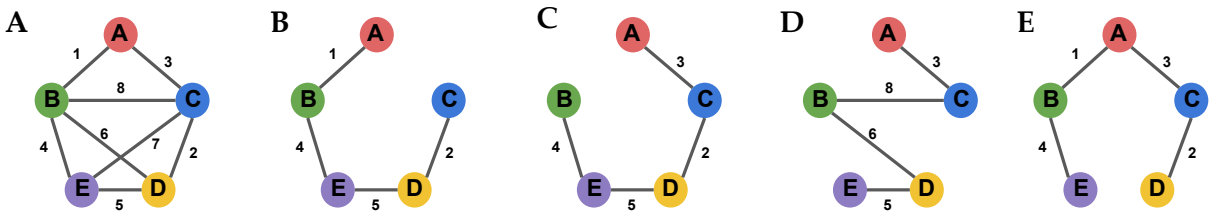


Figure 6.2: (A) Graph $G(V, E, w)$, where V represents vertices, E represents edges and w represents weights of the edges. (B)-(E) Multiple Spanning tree corresponding to the graph G. (E) Minimum Spanning tree for Graph G.

6.2.4 ARCANE-ROG: Algorithm for Reconstruction of Cancer Evolution from Single-Cell Data via Robust Graph Learning

The methodology of the proposed algorithm, ARCANE-ROG, is shown in Figure 6.3. It consists of three steps for identifying clonal evolution from the noisy single-cell data. In the first step, denoised and complete data is recovered using robust graph learning from noisy and incomplete single-cell DNA data. In this step, we also learn an adjacency graph simultaneously. The number of clones is determined using the adjacency matrix in the second step. The final step infers a clonal tree using the clones inferred in step 2 via the minimum spanning tree algorithm.

Step 1: Recover denoised and complete matrix from the noisy incomplete data matrix

The first step of ARCANE-ROG recovers an approximate version of the true genotype matrix from the noisy and incomplete data. The original genotype matrix is a low-rank matrix, where the tumor cells are clustered together into various subclones such that there is little to no variation in the genotype of the cells within the same subclone. Hence, our goal is to recover this low-dimensional subspace of subclones along with the imputation of missing values that are embedded in the noisy and incomplete data matrix. For this task, we have used Robust Graph learning [227] in our work, where the input data is denoised, and an adjacency graph is also learnt simultaneously. Since both these operations are implemented together in a joint framework, they boost each other

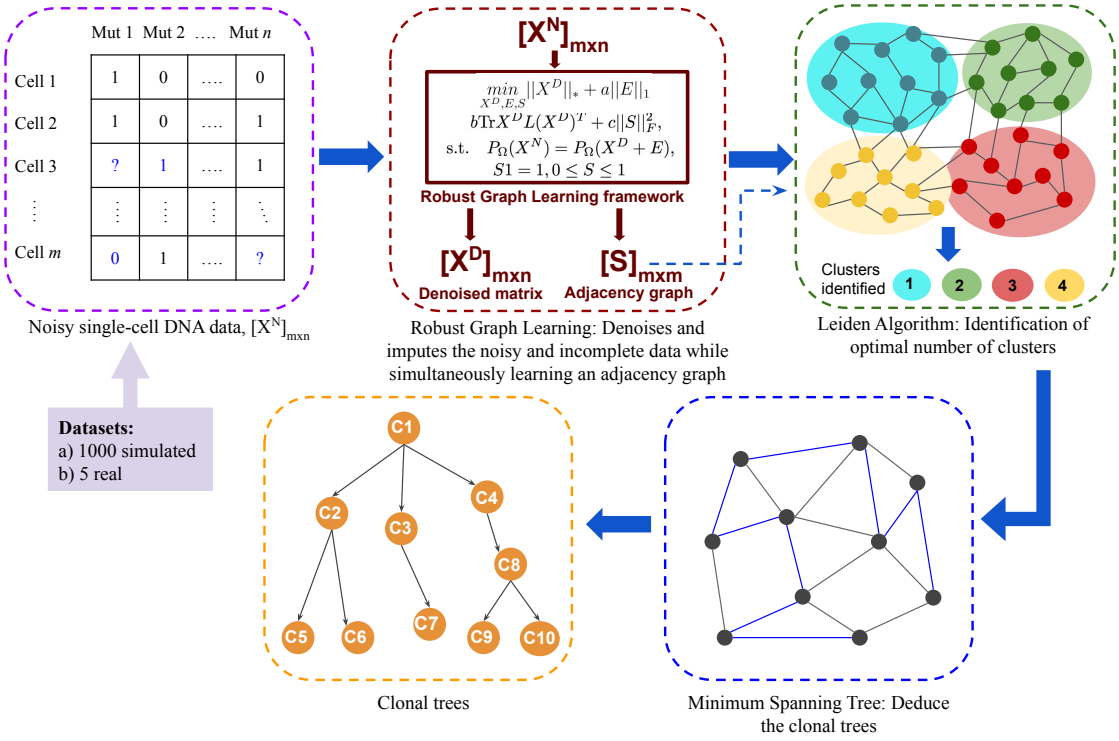


Figure 6.3: Methodology of the proposed ARCANE-ROG method. It consists of three steps for identifying clonal evolution from the noisy single-cell data. In the first step, denoised and complete data is recovered from the noisy and incomplete single-cell DNA data using robust graph learning. In this step, we also learn an adjacency graph simultaneously. In the second step, the number of clones are determined using the adjacency graph. Final step infers clonal tree using the clones inferred in step 2 via minimum spanning tree algorithm.

such that the overall performance of the algorithm is enhanced. Since the robust graph learning algorithm only denoises the noisy data, we extended the existing algorithm to impute missing values in the incomplete input data matrix. The details of this modified algorithm for simultaneous denoising and data imputation are provided in the Appendix C.

Step 2: Identify optimal number of subclones in the data

Once the denoised complete data matrix and similarity matrix are obtained, an optimal number of clusters is identified using the similarity matrix. Similarity matrix is the cell-to-cell adjacency matrix that was learned in the first step. Leiden algorithm is applied on this adjacency graph to infer subclones present in the data.

Step 3: Infer clonal evolution tree from the denoised data and the subclones

Mutations corresponding to each subclone are found using the subclone information and the denoised data, which is referred to as the genotype of the subclones. Euclidean distance between each pair of subclones is calculated based on their subclonal genotypes, and then the minimum spanning tree among the different subclones is inferred via their euclidean distances. The minimum spanning tree so obtained is the clonal tree.

6.2.5 Evaluation Metrics

We have used four different evaluation metrics for performance comparison in our work. For evaluating the performance of the denoising framework, reconstruction error and False Positive to False Negative (FPFN) ratio was computed. For assessing the performance of the clustering and clonal tree inference, tree error and V-measure have been calculated. These metrics are specified below.

Reconstruction Error

Reconstruction error is based on the number of mismatched entries between the denoised matrix and the ground truth matrix. The ground truth matrix is the original matrix on which the error was added. Reconstruction error should be as low as possible.

$$\text{R.E.} = \frac{\text{Number of unequal entries between } X^D \ \& \ X^G}{\text{Size of the matrix}(m \times n)} \tag{6.7}$$

FPFN Ratio

FPFN ratio is the ratio of the total number of false positives and false negatives in the denoised matrix to the total number of false positives and false negatives in the ground truth matrix. It should be as low as possible for the denoised matrix.

$$\text{FPFN ratio} = \frac{\text{FP}(X^D) + \text{FN}(X^D)}{\text{FP}(X^N) + \text{FN}(X^N)} \tag{6.8}$$

Tree Reconstruction Error

Tree reconstruction error, also known as the tree distance error, is calculated as the average differences between the shortest pairwise distance computed for the reconstructed

tree and the ground truth tree. Cells are grouped in different clusters in a tree and distance between two cells is computed as the distance between the clusters to which these cells belong to. Pairwise distance matrices for both the inferred tree and the ground truth tree were computed and the difference in between the two matrices yielded the tree reconstruction error. Its value is 0 when the inferred tree is identical to the ground truth tree. In our work, we have reported normalized values of tree distance error by multiplying the error value with the size of the data matrix.

V-measure

V-measure is a cluster evaluation measure. It basically informs about the goodness of the clusters obtained from any algorithm. It is defined as

$$\text{V-measure} = \frac{(1+\beta) \times \text{homogenity} \times \text{completeness}}{(\beta + \text{homogenity}) + \text{completeness}} \tag{6.9}$$

where β is a factor that provides more weight to either homogeneity or completeness. Homogeneity measures the similarity between the samples in a cluster, while completeness determines if similar samples are grouped in the same cluster or not. It is a reliable measure to ascertain the output of clustering as it does not depend on the number of class labels or number of the clusters. V-measure ranges from 0 to 1, with higher values indicating better performance.

6.2.6 Datasets

We have evaluated the performance of our proposed ARCANE-ROG method on the simulated as well as real datasets. We generated a total of 1000 datasets of different sizes under various settings. We have also generated simulated datasets mimicking real datasets to determine the efficacy of our proposed method. Details of the dataset used are given below.

Simulated Data

Five different groups of datasets, namely S1, S2, S3, S4 and S5, were simulated. There were 200 datasets in each group with varying settings to test the effectiveness of ARCANE-ROG. S1 group was simulated with varying values of α (False positive rate) ranging from 0.001 to 0.01, 0.1 and 0.2. 50 datasets for each value of α were simulated. The number of cells was fixed to 500, mutation sites to 500, the number of clones to 10, β (False-negative rate) to 0.2 and γ (Missing rate) to 0.2. S2 group was

simulated with varying values of β (False-negative rate) ranging from 0.1 to 0.2, 0.3 and 0.4. 50 datasets for each value of β were simulated. The number of cells was fixed to 500, mutation sites to 500, the number of clones to 10, α to 0.01 and γ (Missing rate) to 0.2. S3 group was simulated with varying values of γ (Missing rate) ranging from 0.2 to 0.3, 0.4 and 0.5. The number of cells was fixed to 500, mutation sites to 500, the number of clones to 10, α to 0.01 and β to 0.2. 50 datasets for each value of γ were simulated. S4 group was simulated with a varying number of mutation sites (*n*) ranging from 100 to 500, 1000, 2000. 50 datasets for each value of *n* were simulated. The number of cells was fixed to 500, number of clones to 10, α to 0.01, β to 0.2 and γ to 0.2. S5 group was simulated with varying number of cells (*m*) and clones starting from 100 cells with 10 clones to 500 cells with 20 clones, 1000 cells with 30 clones and 2000 cells with 40 clones. 50 datasets for each value of *m* and the number of clones were simulated. The number of mutation sites was fixed at 500, α at 0.01, β at 0.2 and γ at 0.2.

Real Data

Apart from simulated data, we also tested the performance of ARCANE-ROG on five different real datasets. Real datasets were of different sizes with varying rates of missing entries, false positives, and false negatives. Different real datasets used in the study are explained in detail as follows. JAK2-negative myeloproliferative neoplasm dataset contained 58 cells and 712 mutation sites and was initially studied in [216]. This dataset has a high missing rate of 58%. The binarized matrix used in our study was directly downloaded from [220]. Muscle-invasive bladder transitional cell carcinoma single-cell dataset of 44 cells and 443 mutation sites were initially studied in [235]. A total of 55.2% values were missing in this dataset. We used the data matrix provided with oncoNEM software [220]. Clear-cell renal-cell carcinoma dataset consisting of 17 cells and 35 mutations was initially studied in [212]. The false-positive rate was estimated to be 2.67×10^5, and the false-negative rate was estimated to be 0.1643 by Xu et al. This dataset has 22% missing values. We used the processed data matrix provided with the SCITE software [219]. Estrogen-receptor positive (ER^+) breast cancer dataset consisting of 47 cells and 40 mutations was initially studied in [215]. We used the processed data matrix provided with the SCITE software [219]. High grade serious ovarian cancer (HGSOC) dataset of size 420 cells and 48 mutation sites was studied in [221, 236]. We downloaded the data matrix from [221].

There is no ground truth in real single-cell datasets that can be used to ascertain the accuracy of the results obtained on these real datasets. Hence, we performed an ablation study to determine the fidelity of the results via our proposed method. We simulated datasets imitating the characteristics of real datasets in terms of missing

values, false positives, and false negatives. The rates of missing entries, false positives, and false negatives were reported in previous studies. The size of the simulated datasets was fixed to the size of the real datasets. We first generated 50 datasets with the reported percentage of missing values and then introduced the reported number of false positives in the next 50 datasets and finally added false negatives to the last 50 datasets. Such a type of ablation study helps in deducing the trajectory of the performance of the proposed method when different forms of noise are added to the data. Overall, 150 datasets were generated for the three datasets: JAK2-negative myeloproliferative neoplasm, muscle-invasive bladder transitional cell carcinoma and clear-cell renal-cell carcinoma dataset. The missing rate was unknown for the Estrogen-receptor positive (ER^+) breast cancer dataset of size 47×40. Hence, 50 datasets were simulated with only false positives and false negatives. For the high grade serious ovarian cancer dataset of size 420×48, information on only missing rate was available. Hence, 50 synthetic datasets were generated with only missing entries. We also compared the results obtained on two real datasets with the previous findings. We used TARGET (https://software.broadinstitute.org/cancer/cga/target) database and COSMIC (https://cancer.sanger.ac.uk/actionability/home) database for identifying the actionable genes.

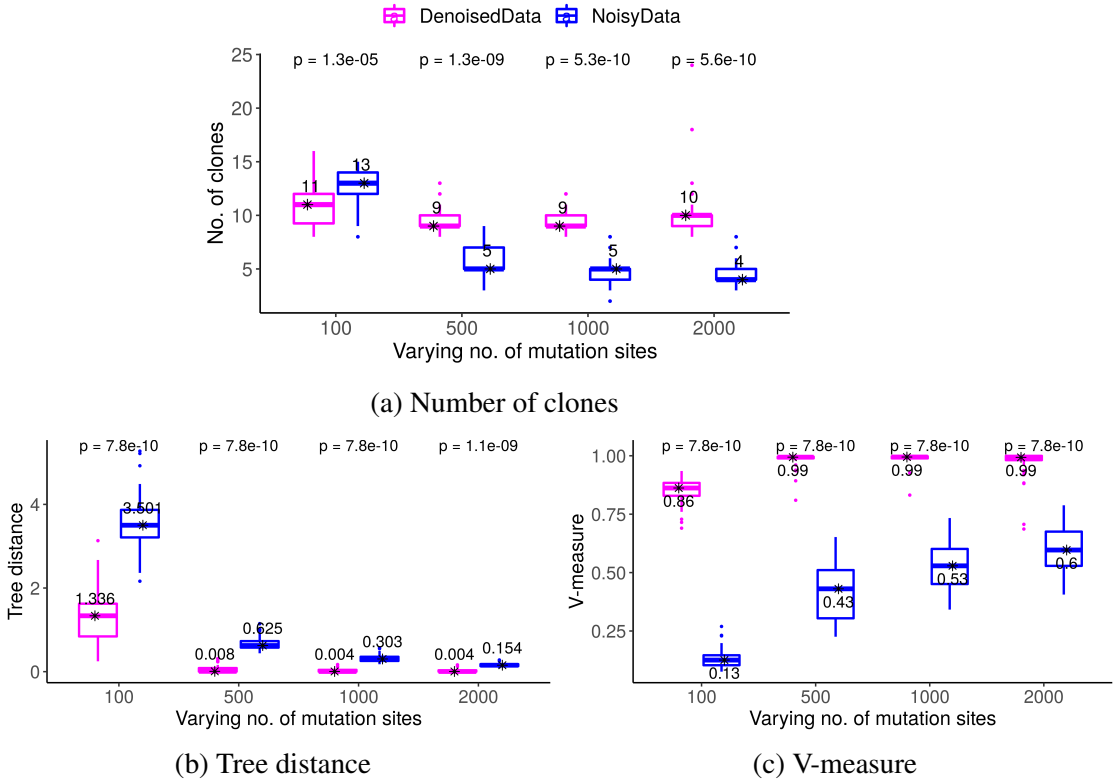
6.3 Results

Our proposed ARCANE-ROG method was compared with the RobustClone [225] method, which is a state-of-the-art method for reconstructing clonal evolution in single-cell data. We have also compared ARCANE-ROG with BnpC [228] and GRMT [229] methods, the results of which have been provided in Appendix 3 Table C.1 and Figures 6.10 and 6.11 respectively. BnpC and GRMT are computationally time-consuming for datasets of size 500×500 and larger. RobustClone, BnpC, and GRMT were run on default parameters.

6.3.1 Impact of denoising in the inference of clonal trajectory

To investigate the role of the denoising process, we performed a few experiments where 50 datasets were generated for each mutation site, i.e. 100, 500, 1000, and 2000 while keeping the number of cells fixed to 500 and the number of clones fixed to 10. α, β and γ were fixed to 0.01, 0.2, 0.2 respectively for the experiments. The number of clones, tree distance and V-measure were calculated on the noisy and denoised simulated data. It is evident from Figure 6.4(a) that for the noisy data, the number of clones was overestimated when the number of mutation sites were 100 and underestimated when the

number of mutation sites was greater than 100. On the contrary, the clones inferred via denoised data were close to 10 at all settings. Similarly, tree distance was high, and V-measure was low for the noisy data compared to the denoised data for all values of mutation sites. It can, thus, be established that noisy data hampers the process of reconstruction of clonal evolution.



(a) Number of clones

(b) Tree distance

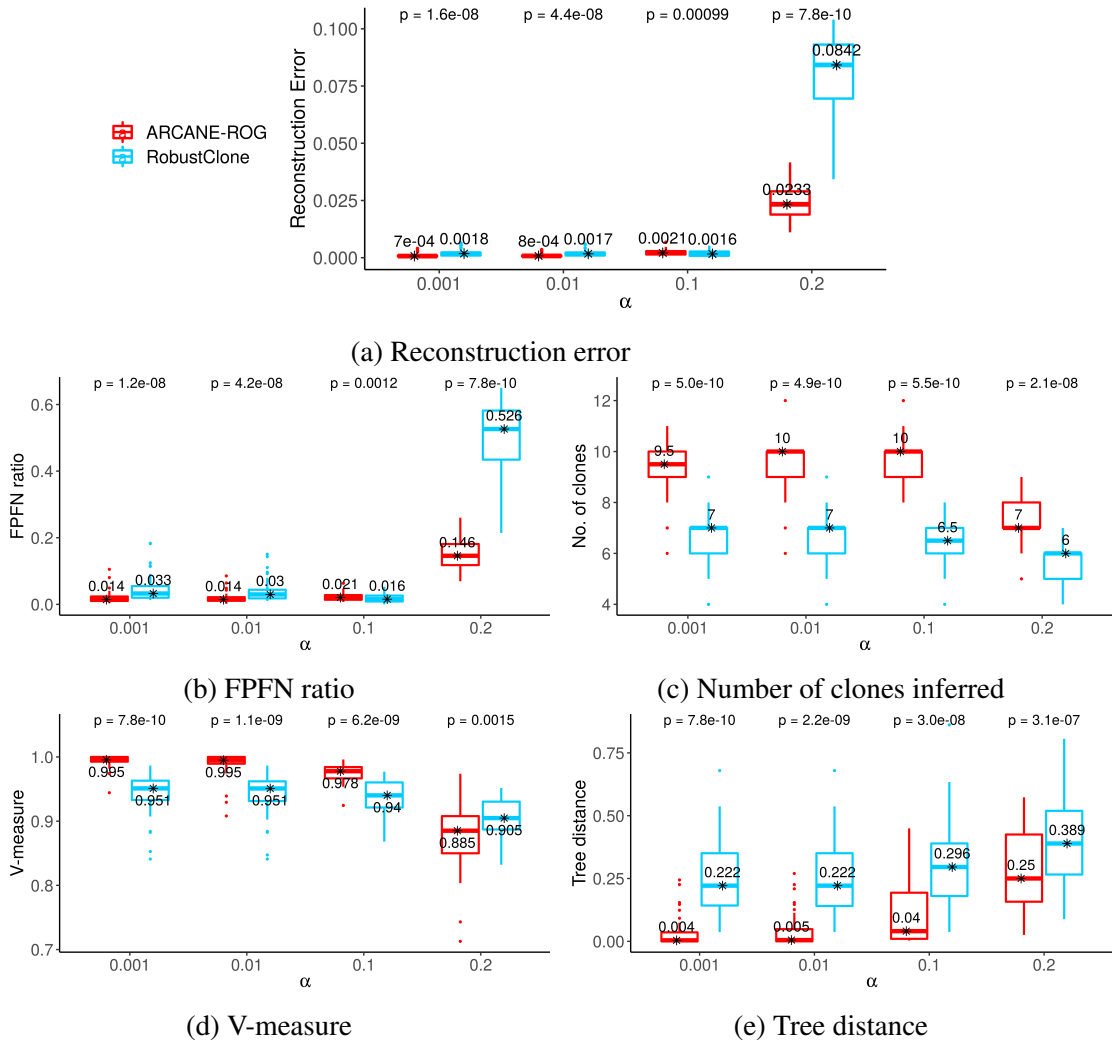
(c) V-measure

Figure 6.4: Boxplots for the comparison of the noisy data with the denoised data for the simulated datasets. (a) The number of clones are overestimated when the number of mutation sites are 100 and underestimated when the number of mutation sites are greater than 100. On the contrary, the number of clone inferred via denoised data are close to 10 at all settings. (b) and (c) Similarly, tree distance was high and V-measure was low for the noisy data as compared to denoised data at all values of mutation sites.

6.3.2 Performance on Simulated Dataset

Performance with varying α

In the S1 dataset, α was varied from 0.001 to 0.01, 0.1 and 0.2, while the number of cells was set to 500, the number of mutation sites to 500, β to 0.2 and γ to 0.2 and the number of clones to 10. With an increase in values of α, reconstruction error increased. Though the difference in the performance was not visible at lower values of alpha, our proposed



(a) Reconstruction error

(b) FPFN ratio

(c) Number of clones inferred

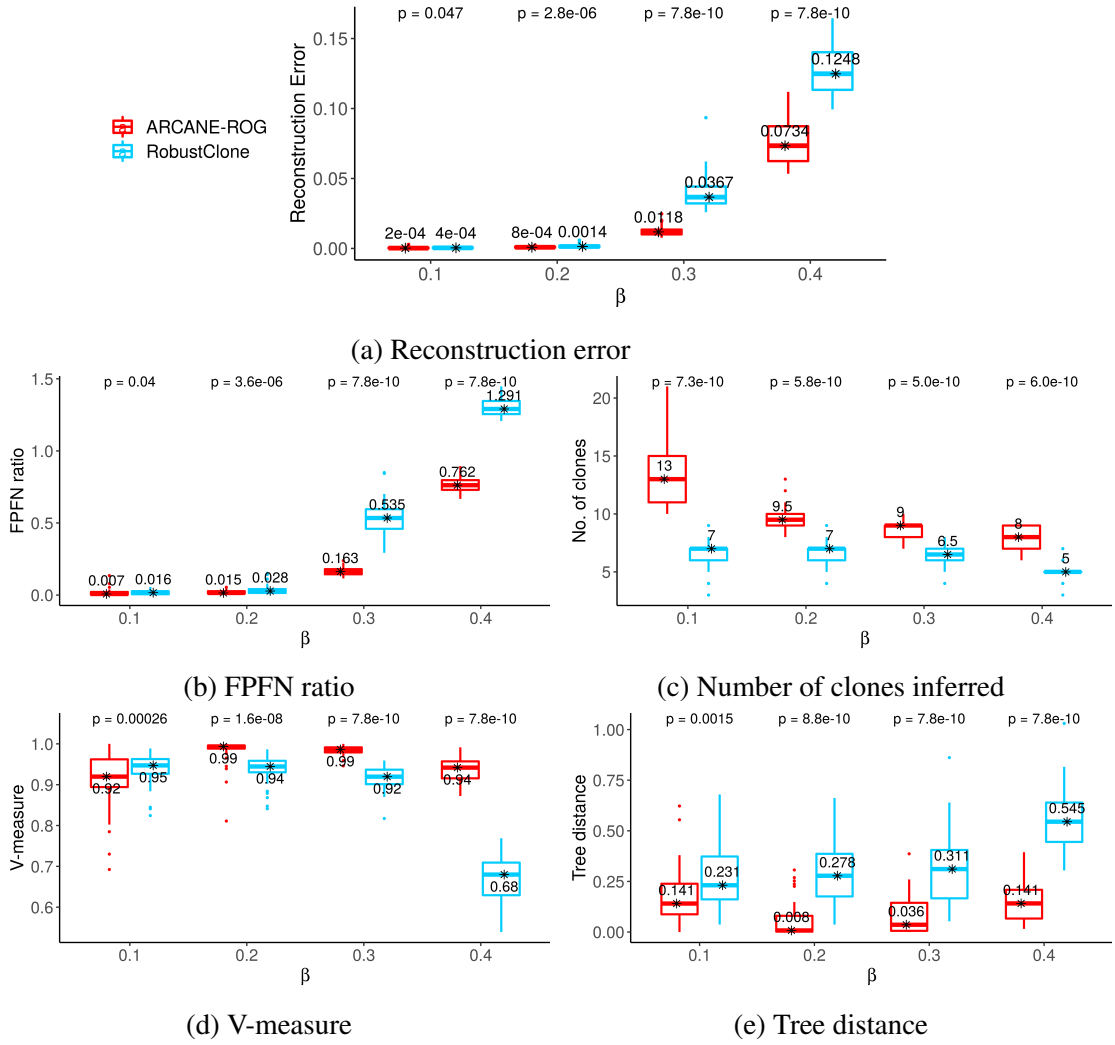
(d) V-measure

(e) Tree distance

Figure 6.5: Boxplots for comparison of the proposed ARCANE-ROG method with RobustClone at varying rates of α. (a) Reconstruction error and (b) FPFN ratio increased slowly with an increase in α but at 0.2, there was a sharp increase leading to the highest value of reconstruction error and FPFN ratio. (c) Number of clones estimated were around 10 at all values of α except at 0.2 where ARCANE-ROG underestimated the number of clones to be around 8. (d) V-measure decreased with an increase in α and was the lowest at 0.2. (e) Tree distance also increased with increase in α and had the maximum value at 0.2 when the number of clones were not inferred accurately. Overall, ARCANE-ROG demonstrated significantly superior performance (p-value < 0.05) as compared to RobustClone at all values of α thereby suggesting that it is robust to false positives.

method performed remarkably better than RobustClone at $\alpha = 0.2$, as shown in the Figures 6.5(a)-6.5(e). Similarly, the FPFN ratio was lower at all varying rates compared to RobustClone. At lower values of α, ARCANE-ROG could perfectly reconstruct the clonal history with nearly zero tree distance. V-measure is also high for ARCANE-ROG compared to RobustClone at all values of α except at 0.2, where the V-measure is slightly less than that of RobustClone. Further, ARCANE-ROG, which uses the Leiden

algorithm, was able to infer the number of clones more accurately than RobustClone. The number of clones inferred via ARCANE-ROG was around 10 for low values of α while for RobustClone, it was around 7. At $\alpha = 0.2$, the range of inferred clones was around 7 for ARCANE-ROG, while that of RobustClone was around 5. Overall, the performance of our proposed method was found to be significantly superior (p-value $<$ 0.05) to the RobustClone with increasing values of false positives.



(a) Reconstruction error

(b) FPFN ratio

(c) Number of clones inferred

(d) V-measure

(e) Tree distance

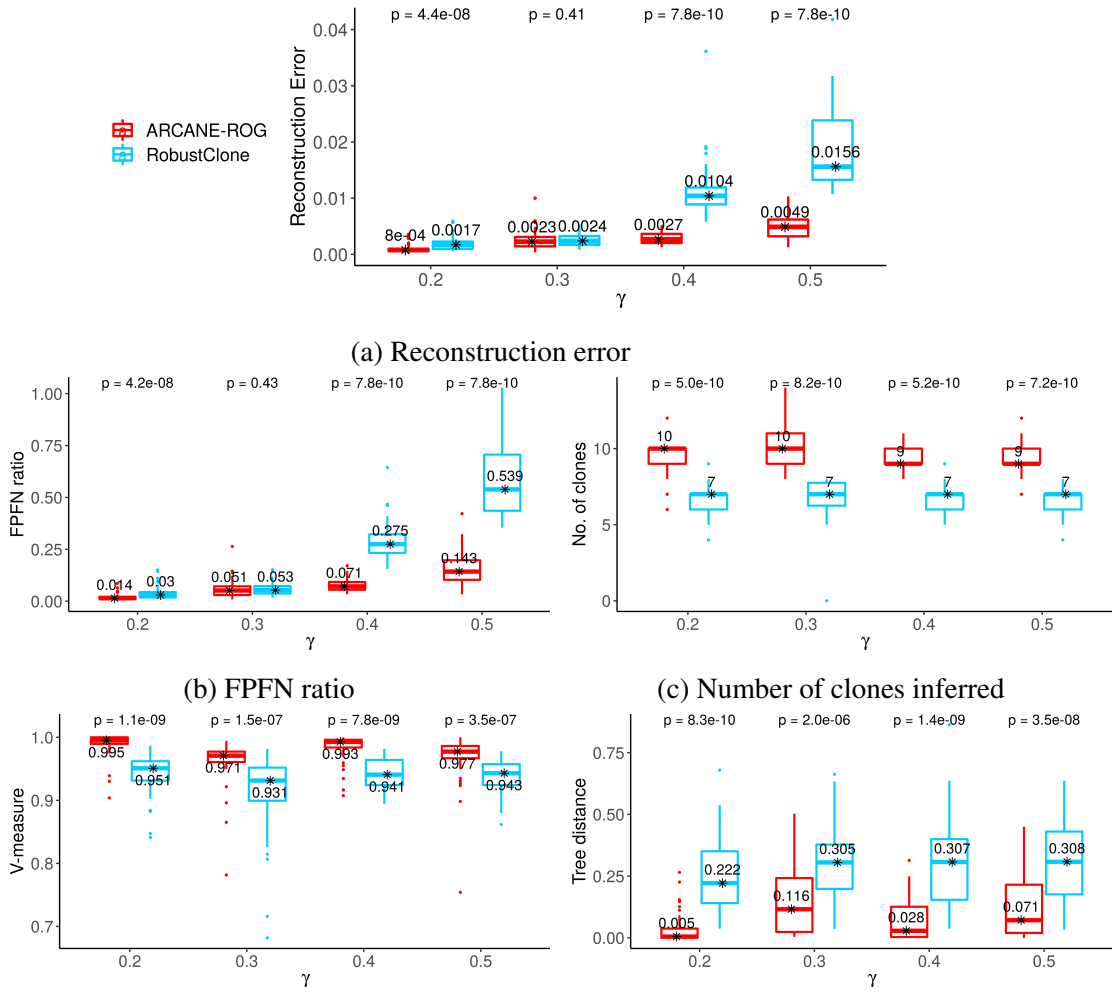
Figure 6.6: Boxplots for comparison of the proposed ARCANE-ROG method with RobustClone at varying rates of β. (a) Reconstruction error and (b) FPFN ratio increased slowly with an increase in β but at 0.2, there was a sharp increase leading to the highest value of reconstruction error and FPFN ratio. (c) Number of clones were overestimated at 0.1 and were around 10 at other values of β. (d) V-measure was low at 0.1 due to overestimation of number of clones. For other values of β, it gradually decreased with an increase in β (e) Tree distance was high owing to overestimation of number of clones at 0.1 and after that, it gradually increased with an increase in β. Overall, the performance of ARCANE-ROG was significantly better (p-value $<$ 0.05) than RobustClone at all values of β thereby suggesting that it is robust to false negatives.

Performance with varying β

In the S2 dataset, β was varied from 0.1 to 0.2, 0.3 and 0.4, while the number of cells was set to 500, the number of mutation sites to 500, α to 0.01 and γ to 0.2 and the number of clones to 10. For $\beta = 0.1$ and 0.2, reconstruction error and FPFN ratio were nearly zero as shown in Figures 6.6(a)-6.6(e). It reveals that the denoised matrix was reconstructed perfectly. Reconstruction error and FPFN ratio gradually increased with an increase in the value of β. However, it remained consistently lower as compared to RobustClone. The gap in the performance became evident at higher levels of β, suggesting the superior performance of our proposed method even at high levels of β. The number of clones was overestimated by ARCANE-ROG, while RobustClone underestimated them at $\beta = 0.1$. Because of overestimation of number of clones, Tree distance and V-measure values were higher for $beta = 0.1$ than for $beta = 0.2, 0.3$. For $\beta = 0.4$, the number of clones inferred was 8, while the ground truth was 10. The inference of the number of clones by ARCANE-ROG was better than that obtained by RobustClone. Owing to better estimation of the number of clones, the tree distance and the V-measure, our proposed method works superior to RobustClone. Overall, ARCANE-ROG performed significantly better (p-value < 0.05) than RobustClone at all values of β.

Performance with varying γ

In the S3 dataset, γ was varied from 0.2 to 0.3, 0.4 and 0.5, while the number of cells was set to 500, the number of mutation sites to 500, α to 0.01 and β to 0.2 and the number of clones to 10. Reconstruction error and FPFN ratio increased with an increase in the unobserved entries. It was lowest when the missing rate was 0.2 and highest when the missing rate was 0.5 as shown in Figures 6.7(a)-6.7(e). There is a significant performance difference between the RobustClone and our proposed method at higher missing rates compared to lower missing rates. The actual number of the clones were set to 10 for the S3 dataset. The number of clones inferred by our proposed method was around 10 at all levels of missing values whereas, for RobustClone, it was close to 7. As the number of clones inferred by ARCANE-ROG was more comparable to ground truth data, tree distance and V-measure were also superior to RobustClone. Lower values of tree distance and high V-measure further suggest that the performance of our proposed method was not much affected by increasing missing rates, and it was able to reconstruct the clonal evolutionary pattern. Overall, ARCANE-ROG performed significantly (p-value < 0.05) superior to the other method at varying rates of missing bases.



(a) Reconstruction error

(b) FPFN ratio

(c) Number of clones inferred

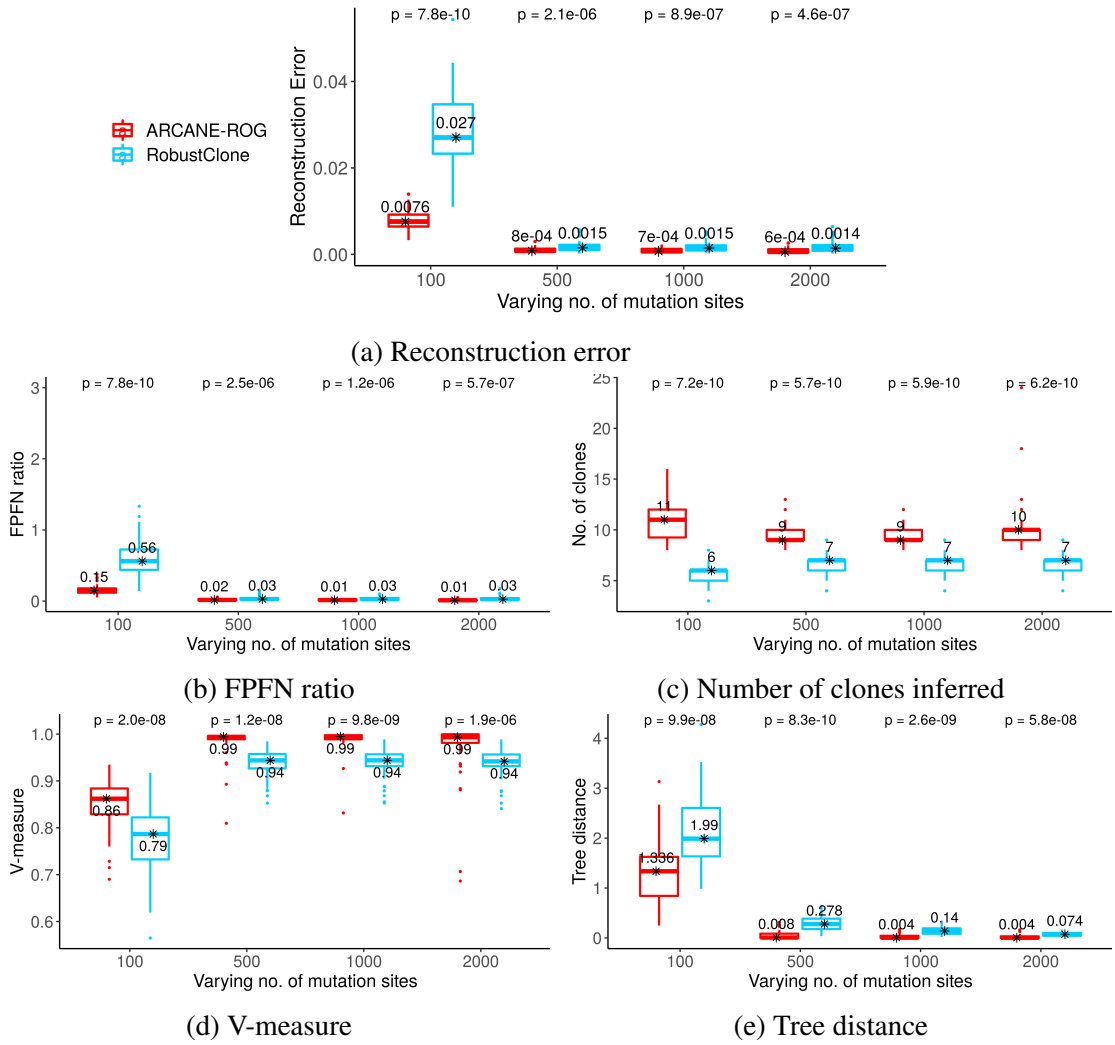
(d) V-measure

(e) Tree distance

Figure 6.7: Boxplots for comparison of the proposed ARCANE-ROG method with RobustClone at varying rates of γ. (a) Reconstruction error and (b) FPFN ratio increased gradually with an increase in γ but had the highest value when the missing rate was 50% i.e. only 50% of the entries are observed in the data. (c) Number of clones were inferred to be around 10 at all missing rates thereby suggesting that the matrices have been accurately reconstructed in the denoising stage. (d) V-measure was nearly 1 at 0.1 (10%) missing rate while it gradually decreased with an increase in the γ. (e) Tree distance was also close to 0 at 0.1(10%) missing rate and after that, it gradually increased with an increase in γ. Overall, ARCANE-ROG significantly (p-value < 0.05) outperformed RobustClone at all percentages of observed values making it robust to the varying rates of missing entries.

Performance with varying mutation sites

In the S4 dataset, the number of mutation sites was varied from 100 to 500, 1000 and 2000, while the number of cells was set to 500, α to 0.01, β to 0.2 and γ to 0.2 and the number of clones to 10. When the number of mutations was 100, the reconstruction error and FPFN ratio was highest, as shown in Figures 6.8(a)-6.8(e). This was because



(a) Reconstruction error

(b) FPFN ratio

(c) Number of clones inferred

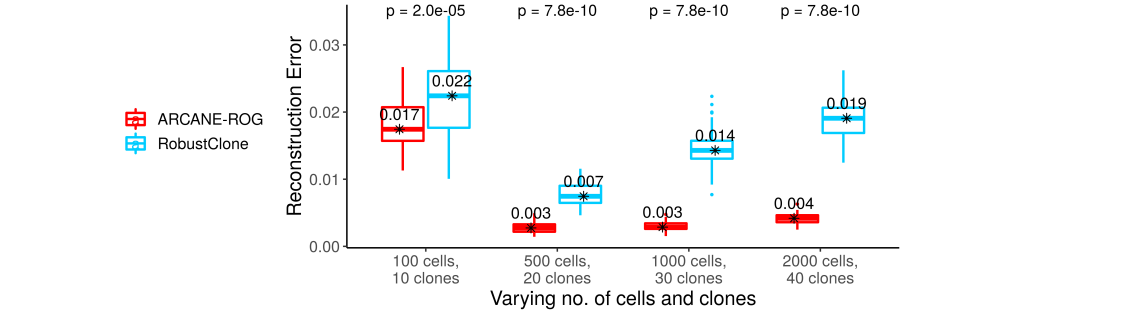
(d) V-measure

(e) Tree distance

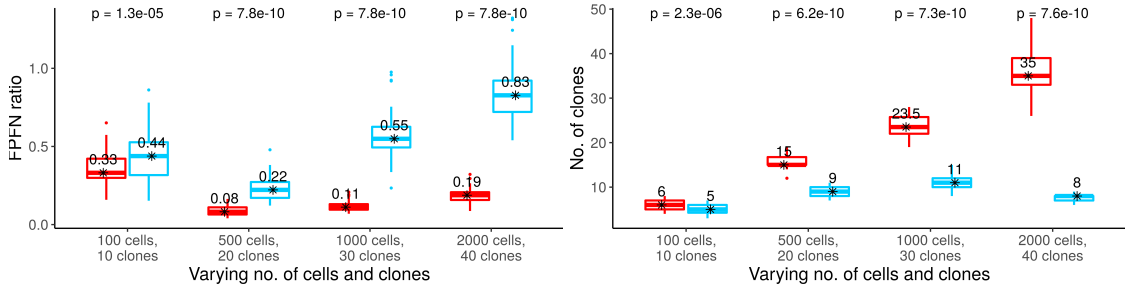
Figure 6.8: Boxplots for comparison of the proposed ARCANE-ROG method with RobustClone at varying number of mutations. (a) Reconstruction error and (b) FPFN ratio had the maximum value when the number of mutations were 100 and it almost decreased to 0 at higher number of mutations. (c) Number of clones were estimated to be around 10 irrespective of the number of mutations. (d) V-measure was the lowest when the number of mutations were 100 while for the rest it was nearly 1. (e) Tree distance gradually decreased with an increase in the number of mutations and was the highest when the number of mutations were 100. Overall, ARCANE-ROG performed significantly better (p-value < 0.05) than RobustClone.

the overall size of the data was small, and hence, the algorithm was not able to recover the original matrix accurately. However, the performance of the proposed method was better than RobustClone. Further, the clones inferred at 100 mutations were above 10 and close to 11. This resulted in high tree distance and low V-measure. However, tree distance and V-measure for ARCANE-ROG were better than RobustClone. When the number of mutation sites changed to 500, 1000 and 2000, reconstruction error and FPFN ratio decreased. The number of clones inferred was also close to 10, with low tree distance and high values of V-measure. Tree distance error was close to 0, and V-

measure was around 0.99 at these settings, indicating the efficacy of ARCANE-ROG. RobustClone also performed comparative to our proposed method, but overall, the performance of ARCANE-ROG significantly surpassed (p-value < 0.05) that of RobustClone.

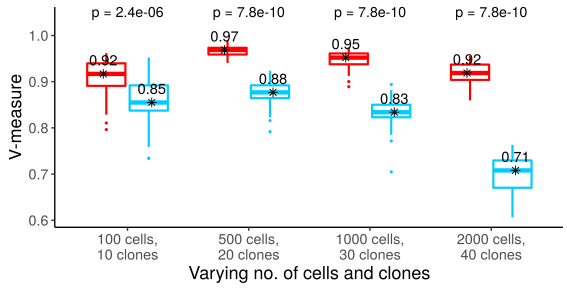


(a) Reconstruction error

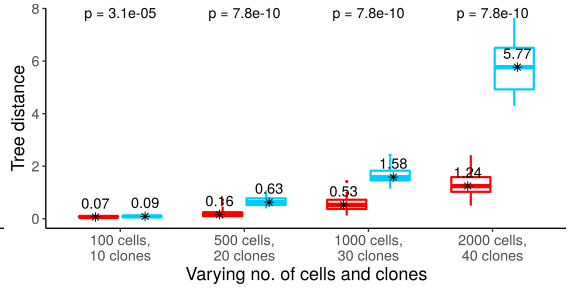


(b) FPFN ratio

(c) Number of clones inferred



(d) V-measure

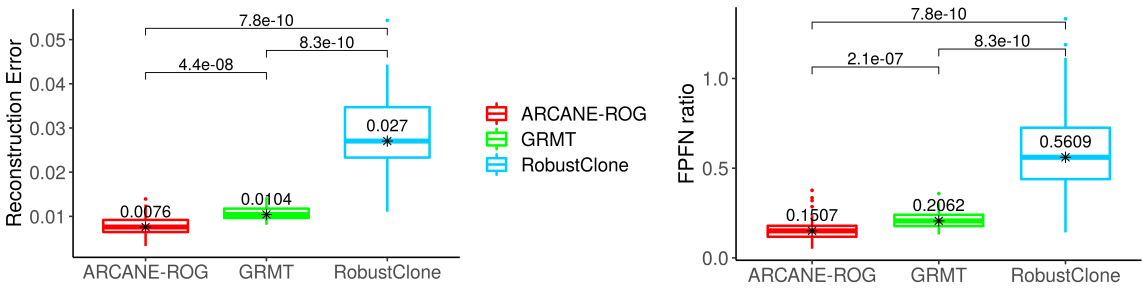


(e) Tree distance

Figure 6.9: Boxplots for comparison of the proposed ARCANE-ROG method with RobustClone at varying number of cells and clones. (a) Reconstruction error and (b) FPFN ratio had the maximum value when the number of cells were 100 and number of clones was set to 10. For 500 cells and 20 clones, reconstruction error and FPFN ratio had the minimum value which gradually increased with an increase in the number of cells and clones. (c) Number of clones estimated by ARCANE-ROG were close to the actual number of clones while RobustClone underestimated the number of clones. (d) V-measure was the lowest when the number of cells and clones were set to 100 and 10 respectively. It had the maximum value for 500 cells and 20 clones after which there was a gradual decrease with an increase in number of cells and clones. (e) Tree distance gradually increased with an increase in the number of cells and clones. Overall, ARCANE-ROG outperformed RobustClone. There was a significant improvement significantly (p-value < 0.05) in the performance.

Performance with varying cells and number of clones

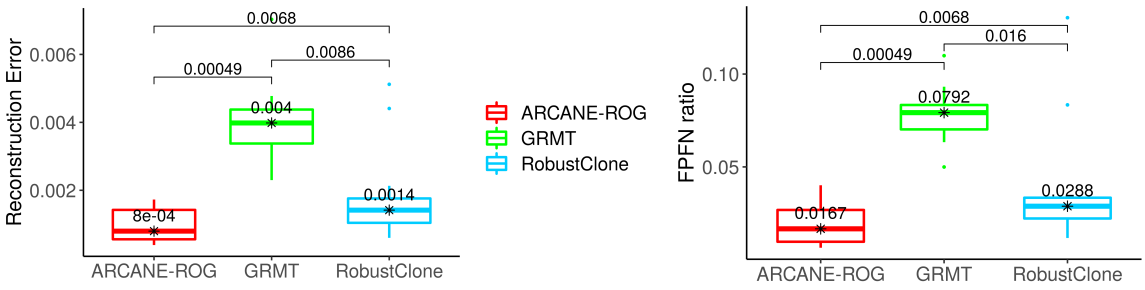
In the S5 dataset, the number of cells was varied along with the number of clones, while the number of mutation sites was set to 500, α to 0.01, β to 0.2 and γ to 0.2. Though the reconstruction error and FPFN ratio were highest at 100 cells and 10 clones in ARCANE-ROG, it was lower than RobustClone. At 500 cells and 20 clones, the reconstruction error and FPFN ratio was lowest in ARCANE-ROG. It increased slightly as the number of cells and the number of clones increased, as shown in Figure 6.9(a)-6.9(e). On the contrary, the performance of RobustClone degraded sharply as the number of cells and the number of clones increased, with the highest reconstruction error and FPFN ratio being observed at 2000 cells and 40 clones. Tree distance increased with an increase in the number of cells and clones in ARCANE-ROG, but it was less than what was observed for RobustClone. The increase in tree distance can also be attributed to the increase in the size of the data matrices. The number of clones inferred by the Leiden method in ARCANE-ROG consistently increased with an increase in the size of the input data and was closer to the ground truth values compared to RobustClone. The performance of RobustClone was poor when the number of cells was 2000, and the number of clones was 40. The number of clones was not inferred correctly by RobustClone. It predicted less number of clones leading to high values of tree distance error and low values of V-measure. Overall, ARCANE-ROG surpassed RobustClone owing to its robust performance. The improvement in the performance was statistically significant (p-value < 0.05).



(a) Reconstruction error

(b) FPFN ratio

Figure 6.10: Boxplots for comparison of the proposed ARCANE-ROG method with GRMT and RobustClone for 50 datasets of size 500×100. (a) Reconstruction error and (b) FPFN ratio had the lowest values for ARCANE-ROG. The values are significantly less as compared to GRMT and RobustClone.



(a) Reconstruction error (b) FPFN ratio

Figure 6.11: Boxplots for comparison of the proposed ARCANE-ROG method with GRMT and RobustClone for 12 datasets of size 500×500. (a) Reconstruction error and (b) FPFN ratio had the lowest values for ARCANE-ROG. The values are significantly less as compared to GRMT and RobustClone.

Comparison with GRMT algorithm

We have compared the results of our method with GRMT. Our method significantly performs better and faster than GRMT. GRMT took more than 24 hours on a 98GB RAM workstation while using 40 cores to compute results on 12 datasets of size 500×500. On the other hand, ARCANE-ROG and RobustClone computed results on the same data within 30 min using single core on the same system. Since GRMT is computationally very heavy, we have computed results on two datasets of small sizes: 500×100 with 500 cells and 100 mutations and 500×500 with 500 cells and 100 mutations. α was set to 0.01, β was set to 0.2, γ was set to 0.2 and number of clones were set at 10 for these two datasets. We could only compute Reconstruction error and FPFN ratio. We could not compute the V-measure and Tree distance for GRMT because GRMT generates tree as output that has a format different from the one used in our method for evaluation. We tried to contact the authors regarding the same because it appears that they modified their tree to calculate the tree evaluation metrics, but we could not get a code for computing tree evaluation metrics from them. Figure 6.10 shows results on dataset of size 500×100. ARCANE-ROG has the lowest median reconstruction error (0.0076) and FPFN ratio (0.1507) and these values were significantly less than those obtained via GRMT (p-value < 0.05) and Robustclone (p-value < 0.05). For this data, GRMT performed significantly better than RobustClone but inferior as compared to ARCANE-ROG. In addition, GRMT was computationally expensive than ARCANE-ROG and RobustClone. It took around 7 hours to compute results on 50 datasets of size 500×100 using 40 cores on a 98GB RAM computer, while ARCANE-ROG and RobustClone computed results within 30 min for 50 datasets using single core on a 98GB RAM computer. Figure 6.11 shows results on dataset of size 500×500. ARCANE-ROG has the lowest median reconstruction error ($8e^{-04}$) and FPFN ratio (0.0167) and

these values were significantly less than those obtained via GRMT (p-value < 0.05) and Robustclone (p-value < 0.05). For this data, GRMT performed significantly inferior to both ARCANE-ROG and RobustClone. In addition, GRMT was computationally expensive than both the methods.

6.3.3 Performance on Real Datasets

There is no ground truth in real single-cell datasets. Therefore, to compare the performance of our proposed method with RobustClone on real datasets, we simulated data imitating the characteristics of real datasets in terms of the missing values, false positives and false negatives. The size of the simulated datasets was fixed to the size of the real datasets.

'JAK2-negative myeloproliferative neoplasm' dataset

Performance of ARCANE-ROG was evaluated for the different datasets generated according to the real data of JAK2-negative myeloproliferative neoplasm. Maximum reconstruction error, tree distance and lowest V-measure was observed when the data was corrupted with missing entries, false positives and false negatives as shown in the Figure C.1. The number of clones inferred by ARCANE-ROG was close to the ground truth in the simulated data, i.e. 5. RobustClone underestimated the number of clones. It can be deduced from these experiments that our proposed method performed superior to RobustClone under varied settings of the simulated dataset. Further, it was able to recover missing entries and remove the noise from the real scSNV data. The adjacency graph learned during the denoising process was used for finding subclones via the Leiden algorithm. 2 clones were identified, and the number of cells in each clone was 29. The pattern of clonal evolution identified in the data was linear, i.e. all the subclones were connected linearly one after the other, as shown in Figure C.6(a). Results deduced via ARCANE-ROG are consistent with the previous findings in [216, 220, 219].

'Muscle-invasive bladder transitional cell carcinoma' dataset

When the simulated data had missing entries, false positives and false negatives, reconstruction error and tree distance were maximum, and V-measure was minimum as shown in the Figure C.2. For all the settings, ARCANE-ROG was able to accurately predict the number of subclones compared to RobustClone. Given the superior performance of ARCANE-ROG at simulated datasets, it can be concluded that it would perform better at the real dataset also. ARCANE-ROG was able to recover missing

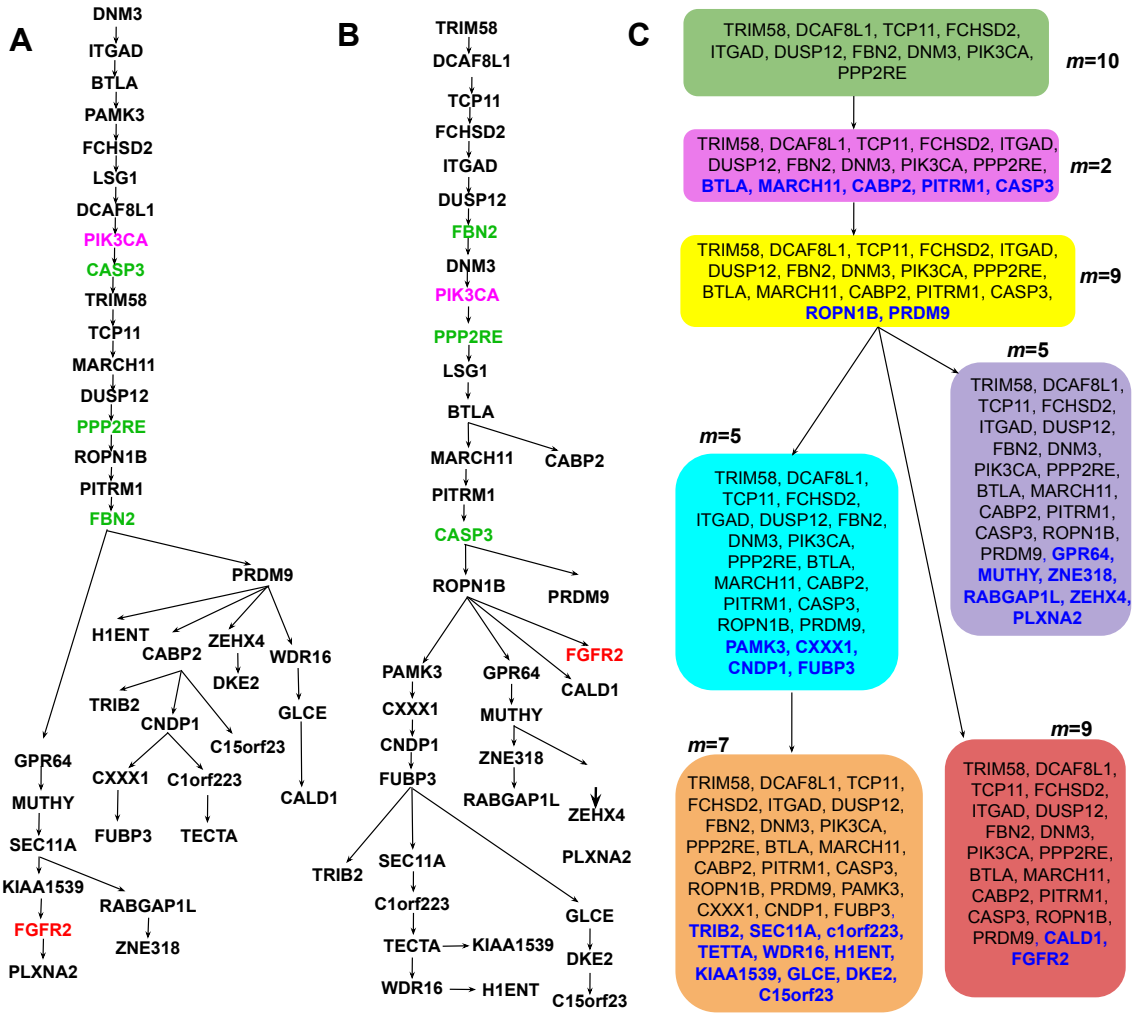


Figure 6.12: Comparison of the results obtained on real dataset of (ER +) breast cancer data. A. MAP (Maximum a posteriori) tree for the cancer data deduced via SCITE. MAP tree provided the order of the mutations acquired progressively during the cancer progression B. Sequential acquisition of mutations in the clonal tree inferred by ARCANE-ROG. Our proposed method provides the order of clones but it does not provide the order of mutations within the clones, hence SCITE was used to infer the sequence of mutations within the clones. Green colored genes indicate non-synonymous mutations in known cancer genes, Magenta colored gene indicate non-synonymous mutations in known cancer genes that are identified actionable according to TARGET and COSMIC database and Red colored genes indicate actionable mutations. C. Clonal tree deduced via ARCANE-ROG. Seven clones were inferred in the data. Blue colored genes indicate the genes acquired in the clone and rest of the genes are carried forward from their parent clone. *m* denotes the number of cells.

values and learn the adjacency graph from the noisy matrix. An adjacency graph was further used to infer three main subclones in the data, with two of the subclones having emerged from the third subclone as shown in Figure C.6(b). There were 16, 17 and 11 cells in each subclone, respectively. The findings were consistent with the conclusions of [235].

'Clear-cell renal-cell carcinoma' dataset

For the simulated data, reconstruction error and tree distance were the maximum when the data was corrupted with missing values, false positives as well false negatives as shown in the Figure C.3. At this setting, the number of clones inferred by both ARCANE-ROG and RobustClone was 2, which is not equal to the ground truth data value of 3. However, V-measure for ARCANE-ROG was still higher than RobustClone, thereby suggesting that ARCANE-ROG could cluster similar cells together efficiently. In other settings, the number of clones estimated by ARCANE-ROG was close to the ground truth value, 3, as opposed to RobustClone, which underestimated the clones. FPFN ratio was calculated only when the simulated data had both the false positives and false negatives along with the missing entries. Overall, ARCANE-ROG outperformed RobustClone. Hence, results inferred by our proposed method can be considered more accurate than RobustClone for the real dataset. Five subclones were identified via our proposed method in the data. The linear pattern was dominant in the clonal tree inferred as shown in Figures C.6(c) and C.7, which was also observed in [215, 219]. FGFR4 was found to be mutated, which is an actionable gene according to the TARGET/COSMIC database.

'Estrogen-receptor positive (ER+) breast cancer' dataset

There were no missing values in the breast cancer dataset; therefore, datasets with only false positives and false negatives were simulated. Our proposed method worked superior as compared to RobustClone in terms of low reconstruction error, low FPFN ratio, low tree distance error and high V-measure as shown in Figure C.4. The number of clones inferred by ARCANE-ROG was close to the ground truth value of 5. Hence, ARCANE-ROG is capable of efficiently identifying the pattern of clonal history. For the real data, seven subclones were inferred, as shown in Figure 6.12. Subclones were arranged in a linear pattern initially and evolved into a more complex branching pattern, later on, giving rise to four different subclones as shown in the Figure C.6(d). These findings were also reported in [212]. Further, in the linear pattern, non-synonymous mutations were identified in genes- *PIK3CA, CASP3, FBN2 and PPP2RE*. PIK3CA and FGFR2 were identified as actionable mutated genes according to the TARGET/COSMIC database. The order of mutations in individual clones was deduced using SCITE [219].

'High grade serious ovarian cancer (HGSOC)' dataset

HGSOC dataset contains 10.7% missing values, and there are no false positives and false negatives in the data. Hence, datasets with only missing entries were generated. ARCANE-ROG had low reconstruction, low tree distance and high V-measure in comparison to another method (Figure C.5). ARCANE-ROG detected the number of clones close to the ground truth value of 8 at all instances. For the real data, our proposed method inferred 6 clones in the data as shown in Figure C.6(e). The results are consistent with the findings in [221]. The number of cells in each cluster are 89 (subclone1), 34 (subclone2), 36 (subclone3), 67 (subclone4), 83 (subclone5) and 111 (subclone6). The root clone has 36 cells which are close to what was predicted in the original study ([221]), where the root clone had 35 cells. Further, the total cells in the two subclones on the left are 123. In the original study by [221], the subclone to the left of the tree had 123 cells. The findings are similar; however, the only difference here is that the clone has been split into two.

6.4 Discussion

In this study, we designed a novel method, ARCANE-ROG, to reconstruct cancer evolutionary patterns from single-cell DNA data. The first step is a graph learning-based framework where the input to the method is a binary genotype matrix containing missing entries and errors in the form of false positives and false negatives. ARCANE-ROG denoises the input matrix and imputes missing entries in the data. It simultaneously learns an adjacency graph during the denoising process. Both the operations are implemented together, resulting in the overall improvement in the algorithm's performance. In the second step, the adjacency graph is used to infer subclones in the data via the Leiden algorithm. Finally, the clonal tree is inferred using a minimum spanning tree algorithm. The performance of the algorithm has been validated on the simulated as well as on the real datasets.

ARCANE-ROG method has been tested on 1000 simulated datasets of different sizes and under different settings via multiple evaluation metrics. Results on all the simulated datasets are compiled and shown in Table 6.1. Reconstruction error and FPFN ratio test the accuracy of the denoising step, V-measure assesses the efficacy of the clustering step and Tree reconstruction error evaluates the correctness of the inferred tree. ARCANE-ROG is robust to changes in the number of cells and clones. It has a statistically significant (p-value < 0.05) low reconstruction error, low tree distance error and low FPFN ratio for small as well as large number of cells as compared to RobustClone. V-measure is also significantly higher than the other method for the varying number of

Table 6.1: Performance comparison of ARCANE-ROG with RobustClone for different values of false positive rate (α), false negative rate (β), missing bases rate (γ), number of mutation sites (n) and number of cells (m) and clones (s). Checkmark indicates that ARCANE-ROG performed significantly better than RobustClone while cross indicates that RobustClone performed better than ARCANE-ROG. RE: Reconstruction error, FPFN ratio: False positive to false negative ratio, TD: Tree distance error and VM: V-measure.

			RE	FPFN ratio	TD	VM	#clones
β= γ = 0.2 *n*= *m*= 500 s= 10	α	0.001	✓	✓	✓	✓	✓
		0.01	✓	✓	✓	✓	✓
		0.1	✗	✗	✓	✓	✓
		0.2	✓	✓	✓	✗	✓
α= 0.01 γ = 0.2 *n*= *m*= 500 s= 10	β	0.1	✓	✓	✓	✓	✓
		0.2	✓	✓	✓	✗	✓
		0.3	✓	✓	✓	✓	✓
		0.4	✓	✓	✓	✓	✓
α= 0.01 β = 0.2 *n*= *m*= 500 s= 10	γ	0.2	✓	✓	✓	✓	✓
		0.3	✓	✓	✓	✓	✓
		0.4	✓	✓	✓	✓	✓
		0.5	✓	✓	✓	✓	✓
α= 0.01 β= γ = 0.2 *m*= 500 s= 10	*n*	100	✓	✓	✓	✓	✓
		500	✓	✓	✓	✓	✓
		1000	✓	✓	✓	✓	✓
		2000	✓	✓	✓	✓	✓
α= 0.01 β= γ = 0.2 *n*= 500	*m*,s	100 cells 10 clones	✓	✓	✓	✓	✓
		500 cells 20 clones	✓	✓	✓	✓	✓
		1000 cells 30 clones	✓	✓	✓	✓	✓
		2000 cells 40 clones	✓	✓	✓	✓	✓

cells. In addition, it is also robust to changes in the number of mutation sites. Though reconstruction is a bit high for the small number of mutation sites, reconstruction error reduces significantly for the larger number of mutations along with low FPFN ratio and low tree distance error. Further, ARCANE-ROG is resilient to alterations in false positives, false negatives and missing bases.

We have further evaluated the performance of ARCANE-ROG on real datasets of varying sizes. Additionally, we have generated simulated datasets imitating characteristics of real data and tested the performance of ARCANE-ROG on these datasets (Table 6.2). ARCANE-ROG can reconstruct clonal patterns very similar to the ground truth clonal pattern of the simulated datasets under diverse settings. The tree distance error, reconstruction error and low FPFN ratio are statistically significantly (p-value < 0.05) lower

Table 6.2: Performance comparison of ARCANE-ROG with RobustClone for simulated datasets generated for real data under varying conditions of false positive rate (α), false negative rate (β), and missing bases rate (γ). Checkmark indicates that ARCANE-ROG performed significantly better than RobustClone while cross indicates that RobustClone performed better than ARCANE-ROG. RE: Reconstruction error, FPFN: False positive to false negative ratio, TD: Tree distance error and VM: V-measure. R1: JAK2-negative myeloproliferative neoplasm, R2:Muscle-invasive bladder transitional cell carcinoma, R3: real Clear-cell renal-cell carcinoma, R4: ER(+) breast cancer and R5: High grade serious ovarian cancer

		RE	FPFN	TD	VM	#clones
R1	γ=0.58	✓	-	✓	✓	✓
	γ=0.58 , α=6.4x10^{-5}	✓	-	✓	✓	✓
	γ=0.58 , α=6.4x10^{-5}, β=0.42	✓	✓	✓	✓	✓
R1	γ=0.55	✓	-	✓	✓	✓
	γ=0.55 , α=6.7x10^{-5}	✓	-	✓	✓	✓
	γ=0.55 , α=6.7x10^{-5}, β=0.4	✓	✓	✓	✓	✓
R1	γ=0.22	✓	-	✓	✓	✓
	γ=0.22 , α=2.67x10^{-5}	✓	-	✓	✓	✓
	γ=0.22 , α=2.67x10^{-5}, β=0.16	✓	✓	✓	✓	✓
R4	α=1.24x10^{-6}, β=0.097	✓	✓	✓	✗	✓
R5	γ=0.11	✓	-	✓	✓	✓

than RobustClone. High values of V-measure indicate that cells have been accurately assigned to different clusters.

6.4.1 Significance of denoising in clonal trajectory inference

Denoising the single-cell data is important because noisy data may impair the inference of clonal trajectory. It is evident from Figure 6.4(a) that in the presence of noise in the input data, the number of clones was overestimated for $n = 100$ and underestimated for $n > 100$. n represents the number of mutation sites. On the contrary, the clones inferred via denoised data were close to 10 at all settings. A similar observation was made for tree distance and V-measure which were high and low respectively for the noisy data compared to the denoised data for all values of n. Hence, it can be corroborated from the experiments that denoising is a critical step for the accurate identification of clones and precise inference of clonal trajectory from the noisy data.

6.4.2 Significance of deducing optimal number of clones and hierarchical order of mutations

ARCANE-ROG can predict an optimal number of clones compared to RobustClone, which underestimates the number of clones. First of all, clones are inferred from the denoised matrix via the Louvain-Jaccard algorithm in RobustClone. Recently, it was shown that the Leiden algorithm outperforms Louvain by determining better partitions in the data [226], contrary to the Louvain method, which may yield poorly connected clusters in the data. Secondly, improper denoising of the noisy matrix may also lead to a sub-optimal number of clusters in the data. However, in ARCANE-ROG, we use an adjacency matrix learnt during the denoising step to infer the optimal number of clones via the Leiden method, which ensures the identification of the optimal number of clones in the data. Consequently, the clonal tree constructed by ARCANE-ROG is more precise and accurate. We validated our findings on real datasets.

As of now, knowledge on the predetermined order of mutations, i.e., the order in which they are acquired in cancer, is limited, but the sequential acquisition of the mutations does influence cancer progression. Thus, deciphering a correct order of mutation may assist us in drawing more relevant and significant biological findings from the data. The mutation tree inferred by ARCANE-ROG for the breast cancer data was compared with the MAP tree inferred by SCITE. The clonal tree initially grouped the cells having similar mutations into clones, and then we deduced the sequence of mutations in individual clones by applying SCITE. Figure 6.12 shows the variation in the sequence of mutations. According to the TARGET and COSMIC database, genes in red color are actionable genes, genes in green indicate non-synonymous mutations in known cancer genes and genes in magenta color are both non-synonymous mutations in known cancer and actionable genes. FBN2 precedes PIK3CA and PPP2RE precedes CASP3 in the tree inferred by ARCANE-ROG (Figure 6.12B) contrary to what is observed in the tree inferred by SCITE in Figure 6.12A. In a recent study ([237]), FAK/ERK signaling pathway was found to be inhibited by the suppression of the FBN2 gene in lung cancer, which validates our finding that the FBN2 gene may mutate before PIK3CA as the FAK signaling pathway affects PIK3 pathway. The progressive acquisition of the mutations is, thus, dependent on the initial clustering of cells. Hence, it is important to deduce the optimal number of clusters from the data as sub-optimal clusters may affect the order of mutations and lead to inaccurate information on the trajectory of mutations.

Cancer is a highly heterogeneous disease consisting of a clonal and sub-clonal population of mutations. Different combinations of chemotherapy drugs are utilized as treatment therapy to prolong the survival of cancer patients, where each drug targets a specific cancer pathway or a clone. A precise and accurate estimation of the clones and

their trajectory pattern may aid in comprehending the phenomenon of drug resistance in patients and assist in deciding the treatment therapy for the patient.

6.4.3 Conclusion

ARCANE-ROG performs well for small datasets of size 17×35 to large datasets of size 2000×500 under varying conditions of false-positive rate, false-negative rate and missing bases, thereby divulging its robustness. Further, ARCANE-ROG can deduce a more precise estimation of the cancer evolutionary pattern from real data. It has significantly outperformed RobustClone in terms of reconstruction error, FPFN ratio, number of clones, tree distance and V-measure. Our proposed method is a reliable and computationally fast method for recovering clonal patterns from single-cell data of all sizes, capable of efficiently dealing with the increasing size of the single-cell data. Overall, the proposed method is an improvement over the existing methods as it enhances cluster assignment and inference on clonal hierarchies. The biological information, thus, deduced would be superior in understanding the sequence of molecular oncogenesis and drug resistance in this era of targeted therapy.

Chapter 7

Conclusion and Future Work

In this dissertation, we proposed robust and efficient solutions to address challenges in cancer genomics. We successfully validated the significance of our proposed methods qualitatively as well as quantitatively.

In chapter 2, we proposed a CS-based framework for addressing the problem of missing values in gene expression data. The novelty of this method is the utilization of row and column sparsity of the gene expression matrix in the Discrete Cosine Transform domain to recover missing values. We demonstrated the robustness of the proposed method on different cancer datasets at low and high observability of data. We further revealed the significance of the imputation via classification and biological pathway analysis. The proposed method was tested mainly on bulk RNA (microarray) data. However, this work could be extended to impute missing values in single-cell RNA sequencing (scRNA) data. Missing values in scRNA arise from dropout events which lead to non biological zero gene expression values and might negatively impact the following data analysis. The method needs to be adjusted to account for the properties of single-cell data. scRNA represents the expression values of individual cells across genes where cells are indicative of different cell types. DCT might not work for this data, so we may need to explore other constraints in addition to nuclear norm for imputing missing values.

In chapter 3, we studied the genomic landscape of clonal evolution in Multiple Myeloma (MM) using the bulk-sequencing whole-exome data of 62 MM patients collected at two time points at AIIMS, New Delhi, first at the time of diagnosis, followed by a second instant on the progression of MM. A comparative analysis of the variants at the two time points along with an in-depth analysis of evolving founder clones revealed multiple driver mutations, including those known to be actionable. Based on these actionable genes, medical treatment tailored to the genetic landscape of the patient could be provided to slow down the progression of disease or to prevent relapse in the patients. Branching evolution was observed in among 72.58% patients, of whom 64.51% had low TMBs and 61.29% had 2 or more founder clones. The hypermutator patients (with high TMB levels 10 to 100) showed a significant decrease in their TMBs from diagnosis to progression. Fall in the subclonal driver mutations was identified recurrently in genes like PABPC1, BRAF, KRAS, CR1, DIS3 and ATM while an analogous rise in driver mutations was observed in KMT2C, FOXD4L1, SP140, NRAS and other genes. The findings of the study are clinically relevant and highlight the importance of

evaluating temporal mutational data for designing better risk stratification strategies and risk adapted combination therapies in future. However, the observations in this data are based on a cohort of 62 patients and in future, these findings could be validated on the a larger cohort of patients. Not only this, single cell exome analysis could be performed on the data to validate the findings of the current study because single cell data provides better resolution. Additionally, whole genome data could be used to further confirm the findings from the exome study as it provides a comprehensive spectrum of mutations. Whole genome data of MM is available through authorized access from dbGaP.

MM is preceded mainly by benign state of Monoclonal Gammopathy of Undetermined Significance (MGUS). MGUS patients do not show any clinical symptoms, unlike MM. However, recent studies have shown that the critical genomic alterations found in MM are also present in MGUS. Thus, in chapter 4, we evaluated the bulk-sequencing exome data of 61 MGUS and 1018 MM patients for a detailed investigation of the change in the mutational spectrum as the disease progresses from MGUS to MM. There was a statistically significant increase in the frequency of all the three categories of variants, non-synonymous (NS), synonymous (SYN), and others (OTH), from MGUS to MM ($P<0.05$). However, there was a statistically significant rise in the TMB values for TMB_NS and TMB_SYN only. It was observed that 3' and 5'UTR mutations were more frequent in MM and might be responsible for driving MGUS to MM via regulatory binding sites. This study also revealed the association of high TMB with newly diagnosed multiple myeloma patients by utilizing the survival data of the MM patients. There was a statistically significant increase in the TMB between patients with poor outcome and superior outcome. A statistically significant association between the APOBEC activity and poor overall survival in MM was discovered. These findings have potential clinical relevance and can assist in designing risk-adapted therapies to inhibit the progression of MGUS to MM and prolong the overall survival in high-risk MM patients. However, this result could be validated further in future using a larger MGUS dataset as the size of the MGUS dataset used in the current study was very small compared to MM. It is challenging to diagnose MGUS owing to the benign condition; hence, genomic data of MGUS is not available easily. Therefore, creating a large cohort by collecting genomic data of MGUS is in itself a research problem. For a clear understanding of the factors responsible for the progression of MGUS to MM, it is important to study larger cohort of MGUS patients in future.

In chapter 5, an ethnicity-aware AI-supported risk staging system, Consensus based Risk Staging system (CRSS), was developed for newly diagnosed multiple myeloma patients. The proposed method is based on easy to acquire clinical parameters like age, albumin, β2M, hemoglobin, eGFR, calcium and the presence of cytogenetic abnormalities in the patient, along with ethnicity information. We validated CRSS on two different datasets belonging to two ethnicities. The performance of CRSS was remarkably bet-

ter compared to the existing risk staging gold standard for Myeloma, i.e. Revised ISS (RISS) in terms of p-values, separation on KM curves, hazard ratios and concordance index. In the future, the proposed risk staging method could be extended by including MM datasets belonging to other ethnicities. Considering the impact of ethnicity on disease biology, utilizing datasets of multiple ethnicities will enhance the robustness and utility of the proposed method. In addition to this, other prognostic factors could be included in the model like therapy given and their response to the treatment. This might help in re-evaluating the risk in the patients after therapy and help in deciding the future course of treatment.

In chapter 6, we devised an optimization-based framework for denoising and imputing noisy and incomplete single-cell data to infer patterns of clonal evolution from the denoised and complete single-cell data. We extensively validated our proposed method on multiple simulated datasets using different evaluation metrics such as reconstruction error, False positive to False Negative (FPFN) ratio, Tree distance and V-measure. We also performed an ablation study on real datasets to examine the performance of our method. Our method infers the number of clones present in the single-cell data and the mutations and cells within each clone; however, we could further extend our method to give the ordering of mutations within the different clones. Such an extension will help find a detailed pattern of how mutations evolve within the tumor cells. Single-cell data also suffers from the problem of doublets. Doublets appear when two (or more) cells are falsely considered to be a part of the same single cell during the time of capturing and processing of single-cell data. So, the method can be further improved to deal with doublets, leading to more accurate identification of clonal evolution patterns in the data.

References

[1] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[2] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, 2009.

[3] Jef D Boeke, George Church, Andrew Hessel, Nancy J Kelley, Adam Arkin, Yizhi Cai, Rob Carlson, Aravinda Chakravarti, Virginia W Cornish, Liam Holt, et al. The genome project-write. *Science*, 353(6295):126–127, 2016.

[4] Genomics England. The 100,000 genomes project. *The*, 100:0–2, 2016.

[5] GenomeAsia100K Consortium et al. The genomeasia 100k project enables genetic discoveries across asia. *Nature*, 576(7785):106, 2019.

[6] Brent Ewing, LaDeana Hillier, Michael C Wendl, and Phil Green. Base-calling of automated sequencer traces usingphred. i. accuracy assessment. *Genome research*, 8(3):175–185, 1998.

[7] Geraldine A Van der Auwera, Mauricio O Carneiro, Christopher Hartl, Ryan Poplin, Guillermo Del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, et al. From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, 43(1):11–10, 2013.

[8] Ravi K Patel and Mukesh Jain. Ngs qc toolkit: a toolkit for quality control of next generation sequencing data. *PloS one*, 7(2):e30619, 2012.

[9] Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data, 2010.

[10] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[11] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011.

[12] Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

[13] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3):1–10, 2009.

[14] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[15] Grzegorz M Boratyn, Jean Thierry-Mieg, Danielle Thierry-Mieg, Ben Busby, and Thomas L Madden. Magic-blast, an accurate rna-seq aligner for long and short reads. *BMC bioinformatics*, 20(1):1–19, 2019.

[16] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[17] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[18] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3):568–576, 2012.

[19] Yu Fan, Liu Xi, Daniel ST Hughes, Jianjun Zhang, Jianhua Zhang, P Andrew Futreal, David A Wheeler, and Wenyi Wang. Muse: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome biology*, 17(1):1–11, 2016.

[20] David Benjamin, Takuto Sato, Kristian Cibulskis, Gad Getz, Chip Stewart, and Lee Lichtenstein. Calling somatic snvs and indels with mutect2. *BioRxiv*, page 861054, 2019.

[21] David E Larson, Christopher C Harris, Ken Chen, Daniel C Koboldt, Travis E Abbott, David J Dooling, Timothy J Ley, Elaine R Mardis, Richard K Wilson, and Li Ding. Somaticsniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317, 2012.

[22] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel. Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, 2012.

[23] Ken Chen, John W Wallis, Michael D McLellan, David E Larson, Joelle M Kalicki, Craig S Pohl, Sean D McGrath, Michael C Wendl, Qunyuan Zhang, Devin P Locke, et al. Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6(9):677–681, 2009.

[24] Kai Ye, Li Guo, Xiaofei Yang, Eric-Wubbo Lamijer, Keiran Raine, and Zemin Ning. Split-read indel and structural variant calling using pindel. In *Copy Number Variants*, pages 95–105. Springer, 2018.

[25] William McLaren, Bethan Pritchard, Daniel Rios, Yuan Chen, Paul Flicek, and Fiona Cunningham. Deriving the consequences of genomic variants with the ensembl api and snp effect predictor. *Bioinformatics*, 26(16):2069–2070, 2010.

[26] Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

[27] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, 2012.

[28] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.

[29] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003.

[30] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4): 248–249, 2010.

[31] Mark F Rogers, Hashem A Shihab, Matthew Mort, David N Cooper, Tom R Gaunt, and Colin Campbell. Fathmm-xf: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, 34(3):511–513, 2018.

[32] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–315, 2014.

[33] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, et al. Cosmic: the catalogue of somatic mutations in cancer. *Nucleic acids research*, 47(D1):D941–D947, 2019.

[34] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44(D1):D862–D868, 2016.

[35] Debyani Chakravarty, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E Rudolph, Rona Yaeger, Tara Soumerai, Moriah H Nissan, et al. Oncokb: a precision oncology knowledge base. *JCO precision oncology*, 1:1–16, 2017.

[36] Qian Xiang, Xianhua Dai, Yangyang Deng, Caisheng He, Jiang Wang, Jihua Feng, and Zhiming Dai. Missing value imputation for microarray gene expression data using histone acetylation information. *BMC bioinformatics*, 9(1):252, 2008.

[37] Ash A Alizadeh, Michael B Eisen, R Eric Davis, Chi Ma, Izidore S Lossos, Andreas Rosenwald, Jennifer C Boldrick, Hajeer Sabet, Truc Tran, Xin Yu, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503, 2000.

[38] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[39] Ki-Yeol Kim, Byoung-Jin Kim, and Gwan-Su Yi. Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC bioinformatics*, 5(1): 160, 2004.

[40] Lígia P Brás and José C Menezes. Improving cluster-based missing value estimation of DNA microarray data. *Biomolecular engineering*, 24(2):273–282, 2007.

[41] Ming Ouyang, William J Welsh, and Panos Georgopoulos. Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, 20(6):917–923, 2004.

[42] Trond Hellem Bø, Bjarte Dysvik, and Inge Jonassen. LSimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic acids research*, 32(3):e34–e34, 2004.

[43] Xiaobai Zhang, Xiaofeng Song, Huinan Wang, and Huanping Zhang. Sequential local least squares imputation estimating missing value of microarray data. *Computers in biology and medicine*, 38(10):1112–1120, 2008.

[44] Zhipeng Cai, Maysam Heydari, and Guohui Lin. Iterated local least squares microarray missing value imputation. *Journal of bioinformatics and computational biology*, 4(05):935–957, 2006.

[45] Dankyu Yoon, Eun-Kyung Lee, and Taesung Park. Robust imputation method for missing values in microarray data. *BMC bioinformatics*, 8(2):S6, 2007.

[46] Xiaobo Zhou, Xiaodong Wang, and Edward R Dougherty. Missing-value estimation using linear and non-linear regression with Bayesian gene selection. *Bioinformatics*, 19(17):2302–2307, 2003.

[47] Muhammad Shoaib B Sehgal, Iqbal Gondal, and Laurence S Dooley. Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data. *Bioinformatics*, 21(10):2417–2423, 2005.

[48] Miew Keen Choong, Maurice Charbit, and Hong Yan. Autoregressive-model-based missing value estimation for DNA microarray time series data. *IEEE Transactions on information technology in biomedicine*, 13(1):131–137, 2009.

[49] Shigeyuki Oba, Masa-aki Sato, Ichiro Takemasa, Morito Monden, Ken-ichi Matsubara, and Shin Ishii. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, 2003.

[50] Rebecka Jörnsten, Hui-Yu Wang, William J Welsh, and Ming Ouyang. DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics*, 21(22):4155–4161, 2005.

[51] Archana Purwar and Sandeep Kumar Singh. Hybrid prediction model with missing value imputation for medical data. *Expert Systems with Applications*, 42(13): 5621–5631, 2015.

[52] Chong He, Changbo Zhao, Guo-Zheng Li, Wei Zhu, William Yang, and Mary Qu Yang. A hybrid iterative approach for microarray missing value estimation. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1350–1352. IEEE, 2016.

[53] Johannes Tuikkala, Laura Elo, Olli S Nevalainen, and Tero Aittokallio. Improving missing value estimation in microarray data with Gene Ontology. *Bioinformatics*, 22(5):566–572, 2005.

[54] Jianjun Hu, Haifeng Li, Michael S Waterman, and Xianghong Jasmine Zhou. Integrative missing value estimation for microarray data. *BMC bioinformatics*, 7 (1):449, 2006.

[55] Yang Yang, Zhuangdi Xu, and Dandan Song. Missing value imputation for microRNA expression data by using a go-based similarity measure. In *BMC bioinformatics*, volume 17, page S10. BioMed Central, 2016.

[56] Francesco Maura, Even H Rustad, Eileen M Boyle, and Gareth J Morgan. Reconstructing the evolutionary history of multiple myeloma. *Best Practice & Research Clinical Haematology*, 33(1):101145, 2020.

[57] Shaji K Kumar and S Vincent Rajkumar. The multiple myelomas—current concepts in cytogenetic classification and therapy. *Nature reviews Clinical oncology*, 15(7):409–421, 2018.

[58] Ola Landgren and Gareth J Morgan. Biologic frontiers in multiple myeloma: from biomarker identification to clinical practice. *Clinical Cancer Research*, 20 (4):804–813, 2014.

[59] Niels van Nieuwenhuijzen, Ingrid Spaan, Reinier Raymakers, and Victor Peperzak. From mgus to multiple myeloma, a paradigm for clonal evolution of premalignant cells. *Cancer research*, 78(10):2449–2456, 2018.

[60] H Kaufmann, J Ackermann, C Baldia, T Nösslinger, R Wieser, S Seidl, V Sagaster, H Gisslinger, U Jäger, M Pfeilstöcker, et al. Both igh translocations and chromosome 13q deletions are early events in monoclonal gammopathy of undetermined significance and do not evolve during transition to multiple myeloma. *Leukemia*, 18(11):1879–1882, 2004.

[61] AM Rajan and SV Rajkumar. Interpretation of cytogenetic results in multiple myeloma for clinical practice. *Blood cancer journal*, 5(10):e365–e365, 2015.

[62] Salomon Manier, Karma Z Salem, Jihye Park, Dan A Landau, Gad Getz, and Irene M Ghobrial. Genomic complexity of multiple myeloma and its clinical implications. *Nature reviews Clinical oncology*, 14(2):100–113, 2017.

[63] Niccolo Bolli, Hervé Avet-Loiseau, David C Wedge, Peter Van Loo, Ludmil B Alexandrov, Inigo Martincorena, Kevin J Dawson, Francesco Iorio, Serena Nik-Zainal, Graham R Bignell, et al. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nature communications*, 5(1):1–13, 2014.

[64] Marta Chesi and P Leif Bergsagel. Molecular pathogenesis of multiple myeloma: basic and clinical updates. *International journal of hematology*, 97(3):313–323, 2013.

[65] Niccolò Bolli, Francesco Maura, Stephane Minvielle, Dominik Gloznik, Raphael Szalat, Anthony Fullam, Inigo Martincorena, Kevin J Dawson, Mehmet Kemal Samur, Jorge Zamora, et al. Genomic patterns of progression in smoldering multiple myeloma. *Nature communications*, 9(1):1–10, 2018.

[66] Maxime Tarabichi, Iñigo Martincorena, Moritz Gerstung, Armand M Leroi, Florian Markowetz, Paul T Spellman, Quaid D Morris, Ole Christian Lingjærde, David C Wedge, and Peter Van Loo. Neutral tumor evolution? *Nature genetics*, 50(12):1630–1633, 2018.

[67] Yusuke Furukawa and Jiro Kikuchi. Molecular basis of clonal evolution in multiple myeloma. *International Journal of Hematology*, pages 1–16, 2020.

[68] John R Jones, Niels Weinhold, Cody Ashby, Brian A Walker, Chris Wardell, Charlotte Pawlyn, Leo Rasche, Lorenzo Melchor, David A Cairns, Walter M Gregory, et al. Clonal evolution in myeloma: the impact of maintenance lenalidomide and depth of response on the genetics and sub-clonal structure of relapsed disease in uniformly treated newly diagnosed patients. *Haematologica*, 104(7): 1440, 2019.

[69] Niels Weinhold, Cody Ashby, Leo Rasche, Shweta S Chavan, Caleb Stein, Owen W Stephens, Ruslana Tytarenko, Michael A Bauer, Tobias Meissner,

Shayu Deshpande, et al. Clonal selection and double-hit events involving tumor suppressor genes underlie relapse in myeloma. *Blood, The Journal of the American Society of Hematology*, 128(13):1735–1744, 2016.

[70] L Melchor, A Brioli, CP Wardell, A Murison, NE Potter, MF Kaiser, RA Fryer, DC Johnson, DB Begum, S Hulkki Wilson, et al. Single-cell genetic analysis reveals the composition of initiating clones and phylogenetic patterns of branching and parallel evolution in myeloma. *Leukemia*, 28(8):1705–1715, 2014.

[71] Jens G Lohr, Petar Stojanov, Scott L Carter, Peter Cruz-Gordillo, Michael S Lawrence, Daniel Auclair, Carrie Sougnez, Birgit Knoechel, Joshua Gould, Gordon Saksena, et al. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer cell*, 25(1):91–101, 2014.

[72] Brian A Walker, Christopher P Wardell, Lorenzo Melchor, Annamaria Brioli, David C Johnson, Martin F Kaiser, Fabio Mirabella, Lucia Lopez-Corral, Sean Humphray, Lisa Murray, et al. Intraclonal heterogeneity is a critical early event in the development of myeloma and precedes the development of clinical symptoms. *Leukemia*, 28(2):384–390, 2014.

[73] Michael A Chapman, Michael S Lawrence, Jonathan J Keats, Kristian Cibulskis, Carrie Sougnez, Anna C Schinzel, Christina L Harview, Jean-Philippe Brunet, Gregory J Ahmann, Mazhar Adli, et al. Initial genome sequencing and analysis of multiple myeloma. *Nature*, 471(7339):467–472, 2011.

[74] Francesco Maura, Niccoló Bolli, Nicos Angelopoulos, Kevin J Dawson, Daniel Leongamornlert, Inigo Martincorena, Thomas J Mitchell, Anthony Fullam, Santiago Gonzalez, Raphael Szalat, et al. Genomic landscape and chronological reconstruction of driver events in multiple myeloma. *Nature communications*, 10(1):1–12, 2019.

[75] Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013.

[76] Mehmet Kemal Samur, Anil Aktas Samur, Mariateresa Fulciniti, Raphael Szalat, Tessa Han, Masood Shammas, Paul Richardson, Florence Magrangeas, Stephane Minvielle, Jill Corre, et al. Genome-wide somatic alterations in multiple myeloma reveal a superior outcome group. *Journal of Clinical Oncology*, 38 (27):3107, 2020.

[77] Jonathan J Keats, Marta Chesi, Jan B Egan, Victoria M Garbitt, Stephen E Palmer, Esteban Braggio, Scott Van Wier, Patrick R Blackburn, Angela S Baker, Angela Dispenzieri, et al. Clonal competition with alternating dominance in multiple myeloma. *Blood*, 120(5):1067–1076, 2012.

[78] Robert A Kyle, Terry M Therneau, S Vincent Rajkumar, Dirk R Larson, Matthew F Plevak, Janice R Offord, Angela Dispenzieri, Jerry A Katzmann, and

L Joseph Melton III. Prevalence of monoclonal gammopathy of undetermined significance. *New England Journal of Medicine*, 354(13):1362–1369, 2006.

[79] Robert A Kyle, Terry M Therneau, S Vincent Rajkumar, Janice R Offord, Dirk R Larson, Matthew F Plevak, and L Joseph Melton III. A long-term study of prognosis in monoclonal gammopathy of undetermined significance. *New England Journal of Medicine*, 346(8):564–569, 2002.

[80] Salomon Manier, Karma Salem, Siobhan V Glavey, Aldo M Roccaro, and Irene M Ghobrial. Genomic aberrations in multiple myeloma. *Plasma Cell Dyscrasias*, pages 23–34, 2016.

[81] Rafael Fonseca, Emily A Blood, Martin M Oken, Robert A Kyle, Gordon W Dewald, Richard J Bailey, Scott A Van Wier, Kimberly J Henderson, James D Hoyer, David Harrington, et al. Myeloma and the t (11; 14)(q13; q32); evidence for a biologically defined unique subset of patients. *Blood, The Journal of the American Society of Hematology*, 99(10):3735–3741, 2002.

[82] Madhav V Dhodapkar. Mgus to myeloma: a mysterious gammopathy of underexplored significance. *Blood, The Journal of the American Society of Hematology*, 128(23):2599–2606, 2016.

[83] Ankit K Dutta, J Lynn Fink, John P Grady, Gareth J Morgan, Charles G Mullighan, Luen B To, Duncan R Hewett, and Andrew CW Zannettino. Subclonal evolution in disease progression from mgus/smm to multiple myeloma is characterised by clonal stability. *Leukemia*, 33(2):457–468, 2019.

[84] Erin D Pleasance, Philip J Stephens, Sarah O'Meara, David J McBride, Alison Meynert, David Jones, Meng-Lay Lin, David Beare, King Wai Lau, Chris Greenman, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, 463(7278):184–190, 2010.

[85] Christopher Greenman, Philip Stephens, Raffaella Smith, Gillian L Dalgliesh, Christopher Hunter, Graham Bignell, Helen Davies, Jon Teague, Adam Butler, Claire Stevens, et al. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158, 2007.

[86] Erin D Pleasance, R Keira Cheetham, Philip J Stephens, David J McBride, Sean J Humphray, Chris D Greenman, Ignacio Varela, Meng-Lay Lin, Gonzalo R Ordóñez, Graham R Bignell, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191–196, 2010.

[87] S Kasar, J Kim, R Improgo, G Tiao, P Polak, N Haradhvala, MS Lawrence, A Kiezun, SM Fernandes, S Bahl, et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nature communications*, 6(1):1–12, 2015.

[88] Serena Nik-Zainal, Ludmil B Alexandrov, David C Wedge, Peter Van Loo, Christopher D Greenman, Keiran Raine, David Jones, Jonathan Hinton, John

Marshall, Lucy A Stebbings, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, 2012.

[89] Helen Davies, Sandro Morganella, Colin A Purdie, Se Jin Jang, Elin Borgen, Hege Russnes, Dominik Glodzik, Xueqing Zou, Alain Viari, Andrea L Richardson, et al. Whole-genome sequencing reveals breast cancers with mismatch repair deficiency. *Cancer research*, 77(18):4755–4762, 2017.

[90] Michael J Fusco, Howard Jack West, and Christine M Walko. Tumor mutation burden and cancer treatment. *JAMA oncology*, 7(2):316–316, 2021.

[91] Cristina Valero, Mark Lee, Douglas Hoen, Jingming Wang, Zaineb Nadeem, Neal Patel, Michael A Postow, Alexander N Shoushtari, George Plitas, Vinod P Balachandran, et al. The association between tumor mutational burden and prognosis is dependent on treatment context. *Nature genetics*, 53(1):11–15, 2021.

[92] Yogita Sharma, Milad Miladi, Sandeep Dukare, Karine Boulay, Maiwen Caudron-Herger, Matthias Groß, Rolf Backofen, and Sven Diederichs. A pan-cancer analysis of synonymous mutations. *Nature communications*, 10(1):1–14, 2019.

[93] Daniel B Goodman, George M Church, and Sriram Kosuri. Causes and effects of n-terminal codon bias in bacterial genes. *Science*, 342(6157):475–479, 2013.

[94] Joanna L Parmley, Jean-Vincent Chamary, and Laurence D Hurst. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Molecular biology and evolution*, 23(2):301–309, 2006.

[95] D Allan Drummond and Claus O Wilke. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2): 341–352, 2008.

[96] Fran Supek, Nives Škunca, Jelena Repar, Kristian Vlahoviček, and Tomislav Šmuc. Translational selection is ubiquitous in prokaryotes. *PLoS genetics*, 6(6): e1001004, 2010.

[97] Rosina Savisaar and Laurence D Hurst. Exonic splice regulation imposes strong selection at synonymous sites. *Genome research*, 28(10):1442–1454, 2018.

[98] Motoo Kimura. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, 267(5608):275–276, 1977.

[99] 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.

[100] 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.

[101] 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.

[102] Angela Rago, Sara Grammatico, Tommaso Za, Anna Levi, Sergio Mecarocci, Agostina Siniscalchi, Luca De Rosa, Stefano Felici, Velia Bongarzoni, Anna Lina Piccioni, et al. Prognostic factors associated with progression of smoldering multiple myeloma to symptomatic form. *Cancer*, 118(22):5544–5549, 2012.

[103] Maximilian Schinke, Gabriele Ihorst, Justus Duyster, Ralph Wäsch, Martin Schumacher, and Monika Engelhardt. Risk of disease recurrence and survival in patients with multiple myeloma: A German Study Group analysis using a conditional survival approach with long-term follow-up of 815 patients. *Cancer*, 126(15):3504–3515, 2020.

[104] N Howlader, M Krapcho, D Miller, K Bishop, CL Kosary, M Yu, J Ruhl, Z Tatalovich, A Mariotto, DR Lewis, et al. Seer cancer statistics review, 1975-2014, based on november 2016 seer data submission, posted to the seer web site. *Bethesda, MD: National Cancer Institute*, 2017.

[105] Sikander Ailawadhi, Ibrahim T Aldoss, Dongyun Yang, Pedram Razavi, Wendy Cozen, Taimur Sher, and Asher Chanan-Khan. Outcome disparities in multiple myeloma: a SEER-based comparative analysis of ethnic subgroups. *British journal of haematology*, 158(1):91–98, 2012.

[106] 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, The Journal of the American Society of Hematology*, 116 (25):5501–5506, 2010.

[107] Luciano J Costa, Ilene K Brill, James Omel, Kelly Godby, Shaji K Kumar, and Elizabeth E Brown. Recent trends in multiple myeloma incidence and survival by age, race, and ethnicity in the united states. *Blood advances*, 1(4):282–287, 2017.

[108] Benjamin A Derman, Jagoda Jasielec, Spencer S Langerman, Wei Zhang, Andrzej J Jakubowiak, and Brian C-H Chiu. Racial differences in treatment and outcomes in multiple myeloma: a multiple myeloma research foundation analysis. *Blood cancer journal*, 10(8):1–7, 2020.

[109] Sarvari V Yellapragada, Nathanael R Fillmore, Anna Frolov, Yang Zhou, Pallavi Dev, Hassan Yameen, Chizoba Ifeorah, Nhan V Do, Mary T Brophy, and Nikhil C Munshi. Vitamin d deficiency predicts for poor overall survival in white but not african american patients with multiple myeloma. *Blood Advances*, 4(8): 1643, 2020.

[110] Dominik D Alexander, Pamela J Mink, Hans-Olov Adami, Philip Cole, Jack S Mandel, Martin M Oken, and Dimitrios Trichopoulos. Multiple myeloma: a review of the epidemiologic literature. *International journal of cancer*, 120(S12): 40–61, 2007.

[111] Benjamin A Logsdon, Andrew J Gentles, Chris P Miller, C Anthony Blau, Pamela S Becker, and Su-In Lee. Sparse expression bases in cancer reveal tumor drivers. *Nucleic acids research*, 43(3):1332–1344, 2015.

[112] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[113] Minoru Kanehisa, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. New approach for understanding genome variations in KEGG. *Nucleic acids research*, 47(D1):D590–D595, 2018.

[114] Minoru Kanehisa. Toward understanding the origin and evolution of cellular organisms. *Protein Science*, 2019.

[115] 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):128, 2013.

[116] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.

[117] Z. Wen, W. Yin, and Y. Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4(4):333–361, 2012.

[118] Z. Kang, C. Peng, and Q. Cheng. Top-N Recommender System via Matrix Completion. In *AAAI*, pages 179–185, 2016.

[119] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust PCA via gradient descent. In *Advances in neural information processing systems*, pages 4152–4160, 2016.

[120] Arnav Kapur, Kshitij Marwah, and Gil Alterovitz. Gene expression prediction using low-rank matrix completion. *BMC bioinformatics*, 17(1):243, 2016.

[121] Derek L Stirewalt, Soheil Meshinchi, Kenneth J Kopecky, Wenhong Fan, Era L Pogosova-Agadjanyan, Julia H Engel, Michelle R Cronk, Kathleen Shannon Dorcy, Amy R McQuary, David Hockenbery, et al. Identification of genes with abnormal expression changes in acute myeloid leukemia. *Genes, Chromosomes and Cancer*, 47(1):8–20, 2008.

[122] Lucía López-Corral, Luis Antonio Corchete, María Eugenia Sarasquete, María Victoria Mateos, Ramón García-Sanz, Encarna Fermiñán, Juan-José Lahuerta, Joan Bladé, Albert Oriol, Ana Isabel Teruel, et al. Transcriptome analysis reveals molecular profiles associated with evolving steps of monoclonal gammopathies. *Haematologica*, 99(8):1365–1372, 2014.

[123] A. Gupta, S.D. Joshi, and P. Singh. On the approximate discrete KLT of fractional Brownian motion and applications. *Journal of the Franklin Institute*, 355(17):8989–9016, 2018. ISSN 0016-0032. doi: https://doi.org/10.1016/j.jfranklin.2018.09.023. URL `http://www.sciencedirect.com/science/article/pii/S0016003218305970`.

[124] E. Van den Berg and M. P. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912, 2008. doi: 10.1137/080714488. URL `http://link.aip.org/link/?SCE/31/890`.

[125] E. Van den Berg and M. P. Friedlander. SPGL1: A solver for large-scale sparse reconstruction, June 2007. http://www.cs.ubc.ca/labs/scl/spgl1.

[126] J. Cai, E.J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[127] Lili Wang, Alex K Shalek, Mike Lawrence, Ruihua Ding, Jellert T Gaublomme, Nathalie Pochet, Petar Stojanov, Carrie Sougnez, Sachet A Shukla, Kristen E Stevenson, et al. Somatic mutation as a mechanism of Wnt/β-catenin pathway activation in CLL. *Blood*, 124(7):1089–1098, 2014.

[128] T Zhan, N Rindtorff, and Michael Boutros. Wnt signaling in cancer. *Oncogene*, 36(11):1461, 2017.

[129] Filomena De Falco, Rita Sabatini, Beatrice Del Papa, Franca Falzetti, Mauro Di Ianni, Paolo Sportoletti, Stefano Baldoni, Isabella Screpanti, Pierfrancesco Marconi, and Emanuela Rosati. Notch signaling sustains the expression of Mcl-1 and the activity of eIF4E to promote cell survival in CLL. *Oncotarget*, 6(18):16559, 2015.

[130] RG Wickremasinghe, AG Prentice, and AJ Steele. p53 and notch signaling in chronic lymphocytic leukemia: clues to identifying novel therapeutic strategies. *Leukemia*, 25(9):1400, 2011.

[131] Ting-lei Gu, Julie Nardone, Yi Wang, Marc Loriaux, Judit Villén, Sean Beausoleil, Meghan Tucker, Jon Kornhauser, Jianmin Ren, Joan MacNeill, et al. Survey of activated FLT3 signaling in leukemia. *PLoS One*, 6(4):e19169, 2011.

[132] Jing Xu, Nicole Pfarr, Volker Endris, Elias K Mai, NH Md Hanafiah, Nicola Lehners, Roland Penzel, Wilko Weichert, Anthony D Ho, Peter Schirmacher, et al. Molecular signaling in multiple myeloma: association of RAS/RAF mutations and MEK/ERK pathway activation. *Oncogenesis*, 6(5):e337, 2017.

[133] Sangtae Kim, Konrad Scheffler, Aaron L Halpern, Mitchell A Bekritsky, Eunho Noh, Morten Källberg, Xiaoyu Chen, Yeonbin Kim, Doruk Beyter, Peter Krusche, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nature methods*, 15(8):591–594, 2018.

[134] David E Larson, Travis E Abbott, and Richard K Wilson. Using somaticsniper to detect somatic single nucleotide variants. *Current protocols in bioinformatics*, 45(1):15–5, 2014.

[135] Colby Chiang, Ryan M Layer, Gregory G Faust, Michael R Lindberg, David B Rose, Erik P Garrison, Gabor T Marth, Aaron R Quinlan, and Ira M Hall. Speedseq: ultra-fast personal genome analysis and interpretation. *Nature methods*, 12 (10):966–968, 2015.

[136] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Science signaling*, 6(269):pl1–pl1, 2013.

[137] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, 2012.

[138] Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, David Tamborero, Michael P Schroeder, Alba Jene-Sanz, Alberto Santos, and Nuria Lopez-Bigas. Intogen-mutations identifies cancer drivers across tumor types. *Nature methods*, 10(11):1081–1082, 2013.

[139] Anna Schuh, Helene Dreau, Samantha JL Knight, Kate Ridout, Tuba Mizani, Dimitris Vavoulis, Richard Colling, Pavlos Antoniou, Erika M Kvikstad, Melissa M Pentony, et al. Clinically actionable mutation profiles in patients with cancer identified by whole-genome sequencing. *Molecular Case Studies*, 4(2): a002279, 2018.

[140] Natalie Galanina, Rafael Bejar, Michael Choi, Aaron Goodman, Matthew Wieduwilt, Carolyn Mulroney, Lisa Kim, Huwate Yeerna, Pablo Tamayo, Jo-Anne Vergilio, et al. Comprehensive genomic profiling reveals diverse but actionable molecular portfolios across hematologic malignancies: implications for next generation clinical trials. *Cancers*, 11(1):11, 2018.

[141] Paul Deveau, Leo Colmet Daage, Derek Oldridge, Virginie Bernard, Angela Bellini, Mathieu Chicard, Nathalie Clement, Eve Lapouble, Valerie Combaret, Anne Boland, et al. QuantumClone: clonal assessment of functional mutations in cancer based on a genotype-aware method for clonal reconstruction. *Bioinformatics*, 34(11):1808–1816, 2018.

[142] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P

Shah. PyClone: statistical inference of clonal population structure in cancer. *Nature methods*, 11(4):396–398, 2014.

[143] Christopher A Miller, Joshua McMichael, Ha X Dang, Christopher A Maher, Li Ding, Timothy J Ley, Elaine R Mardis, and Richard K Wilson. Visualizing tumor evolution with the fishplot package for r. *BMC genomics*, 17(1):880, 2016.

[144] Nicola D Roberts, R Daniel Kortschak, Wendy T Parker, Andreas W Schreiber, Susan Branford, Hamish S Scott, Garique Glonek, and David L Adelson. A comparative analysis of algorithms for somatic snv detection in cancer. *Bioinformatics*, 29(18):2223–2230, 2013.

[145] Daniel C Koboldt. Best practices for variant calling in clinical sequencing. *Genome Medicine*, 12(1):1–13, 2020.

[146] Li Tai Fang, Pegah Tootoonchi Afshar, Aparna Chhibber, Marghoob Mohiyuddin, Yu Fan, John C Mu, Greg Gibeling, Sharon Barr, Narges Bani Asadi, Mark B Gerstein, et al. An ensemble approach to accurately detect somatic mutations using somaticseq. *Genome biology*, 16(1):1–13, 2015.

[147] Irantzu Anzar, Angelina Sverchkova, Richard Stratford, and Trevor Clancy. Neomutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. *BMC medical genomics*, 12(1):1–14, 2019.

[148] Daniel Auclair, Kenneth Carl Anderson, David Avigan, Giada Bianchi, Noa Biran, Maria Chaudhry, Hearn J Cho, Maggie Furlong, Craig C Hofmeister, Ankit J Kansagra, et al. The myeloma-developing regimens using genomics (mydrug) master protocol., 2019.

[149] Nicholas McGranahan, Andrew JS Furness, Rachel Rosenthal, Sofie Ramskov, Rikke Lyngaa, Sunil Kumar Saini, Mariam Jamal-Hanjani, Gareth A Wilson, Nicolai J Birkbak, Crispin T Hiley, et al. Clonal neoantigens elicit t cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*, 351 (6280):1463–1469, 2016.

[150] Naiyer A Rizvi, Matthew D Hellmann, Alexandra Snyder, Pia Kvistborg, Vladimir Makarov, Jonathan J Havel, William Lee, Jianda Yuan, Phillip Wong, Teresa S Ho, et al. Mutational landscape determines sensitivity to pd-1 blockade in non–small cell lung cancer. *Science*, 348(6230):124–128, 2015.

[151] A Miller, Y Asmann, L Cattaneo, E Braggio, J Keats, D Auclair, S Lonial, SJ Russell, and AK Stewart. High somatic mutation and neoantigen burden are correlated with decreased progression-free survival in multiple myeloma. *Blood cancer journal*, 7(9):e612–e612, 2017.

[152] Denise K Walters, Xiaosheng Wu, Renee C Tschumper, Bonnie K Arendt, Paul M Huddleston, Kimberly J Henderson, Angela Dispenzieri, and Diane F Jelinek. Evidence for ongoing dna damage in multiple myeloma cells as revealed by constitutive phosphorylation of h2ax. *Leukemia*, 25(8):1344–1353, 2011.

[153] Alboukadel Kassambara, Claire Gourzones-Dmitriev, Surinder Sahota, Thierry Rème, Jérôme Moreaux, Hartmut Goldschmidt, Angelos Constantinou, Philippe Pasero, Dirk Hose, and Bernard Klein. A dna repair pathway score predicts survival in human multiple myeloma: the potential for therapeutic strategy. *Oncotarget*, 5(9):2487, 2014.

[154] Francesca Cottini, Teru Hideshima, Rikio Suzuki, Yu-Tzu Tai, Giampaolo Bianchini, Paul G Richardson, Kenneth C Anderson, and Giovanni Tonon. Synthetic lethal approaches exploiting dna damage in aggressive myeloma. *Cancer discovery*, 5(9):972–987, 2015.

[155] Raphaël Szalat, Mehmet Kemal Samur, Mariateresa Fulciniti, Michael Lopez, Puru Nanjappa, Alice Cleynen, Kenneth Wen, Subodh Kumar, Tommaso Perini, Anne S Calkins, et al. Nucleotide excision repair is a potential therapeutic target in multiple myeloma. *Leukemia*, 32(1):111–119, 2018.

[156] Claire Gourzones, Caroline Bret, and Jerome Moreaux. Treatment may be harmful: mechanisms/prediction/prevention of drug-induced dna damage and repair in multiple myeloma. *Frontiers in Genetics*, page 861, 2019.

[157] Dharminder Chauhan, Arghya Ray, Kristina Viktorsson, Jack Spira, Claudia Paba-Prada, Nikhil Munshi, Paul Richardson, Rolf Lewensohn, and Kenneth C Anderson. In vitro and in vivo antitumor activity of a novel alkylating agent, melphalan-flufenamide, against multiple myeloma cells. *Clinical Cancer Research*, 19(11):3019–3031, 2013.

[158] Liangning Hu, Bo Li, Gege Chen, Dongliang Song, Zhijian Xu, Lu Gao, Mengyu Xi, Jinfeng Zhou, Liping Li, Hui Zhang, et al. A novel m phase blocker, dcz3301 enhances the sensitivity of bortezomib in resistant multiple myeloma through dna damage and mitotic catastrophe. *Journal of Experimental & Clinical Cancer Research*, 39(1):1–14, 2020.

[159] Paola Neri, Li Ren, Kathy Gratton, Erin Stebner, Jordan Johnson, Alexander Klimowicz, Peter Duggan, Pierfrancesco Tassone, Adnan Mansoor, Douglas A Stewart, et al. Bortezomib-induced "brcaness" sensitizes multiple myeloma cells to parp inhibitors. *Blood, The Journal of the American Society of Hematology*, 118(24):6368–6379, 2011.

[160] Rumi Ino, Takayuki Saitoh, Yuya Kitamura, Kazuki Homma, Noriyuki Takahashi, Gotoh Nanami, Tetsuhiro Kasamatsu, Hiroaki Shimizu, Makiko Takizawa, Morio Matsumoto, et al. The role and therapeutic target of base excision repair genes in multiple myeloma (mm). *Blood*, 130:4403, 2017.

[161] SM Ashiqul Islam, Yang Wu, Marcos Díaz-Gay, Erik N Bergstrom, Yudou He, Mark Barnes, Mike Vella, Jingwei Wang, Jon W Teague, Peter Clapham, et al. Uncovering novel mutational signatures by de novo extraction with sigprofilerextractor. *BioRxiv*, pages 2020–12, 2021.

[162] Ludmil B Alexandrov, Jaegil Kim, Nicholas J Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, Kyle R Covington, Dmitry A Gordenin, Erik N Bergstrom, et al. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, 2020.

[163] Akanksha Farswan, Lingaraja Jena, Gurvinder Kaur, Anubha Gupta, Ritu Gupta, Lata Rani, Atul Sharma, and Lalit Kumar. Branching clonal evolution patterns predominate mutational landscape in multiple myeloma. *American journal of cancer research*, 11(11):5659, 2021.

[164] Soo-Heang Eo, Hyo Jeong Kang, Seung-Mo Hong, and HyungJun Cho. K-adaptive partitioning for survival data, with an application to cancer staging. *arXiv preprint arXiv:1306.4615*, 2013.

[165] Jan Budczies, Frederick Klauschen, Bruno V Sinn, Balázs Győrffy, Wolfgang D Schmitt, Silvia Darb-Esfahani, and Carsten Denkert. Cutoff finder: a comprehensive and straightforward web application enabling rapid biomarker cutoff optimization. *PloS one*, 7(12):e51862, 2012.

[166] Hiroyuki Yamazaki, Kotaro Shirakawa, Tadahiko Matsumoto, Shigeki Hirabayashi, Yasuhiro Murakawa, Masayuki Kobayashi, Anamaria Daniela Sarca, Yasuhiro Kazuma, Hiroyuki Matsui, Wataru Maruyama, et al. Endogenous apobec3b overexpression constitutively generates dna substitutions and deletions in myeloma cells. *Scientific reports*, 9(1):1–14, 2019.

[167] Emily K Law, Anieta M Sieuwerts, Kelly LaPara, Brandon Leonard, Gabriel J Starrett, Amy M Molan, Nuri A Temiz, Rachel Isaksson Vogel, Marion E Meijer-van Gelder, Fred CGJ Sweep, et al. The dna cytosine deaminase apobec3b promotes tamoxifen resistance in er-positive breast cancer. *Science advances*, 2(10): e1601737, 2016.

[168] Shumei Yan, Fan He, Bei Gao, Huini Wu, Mei Li, Liyun Huang, Jianzhong Liang, Qiuliang Wu, and Yong Li. Increased apobec3b predicts worse outcomes in lung cancer: a comprehensive retrospective study. *Journal of Cancer*, 7(6): 618, 2016.

[169] Yan Du, Xiang Tao, Jing Wu, Huandi Yu, Yinhua Yu, and Hongbo Zhao. Apobec3b up-regulation independently predicts ovarian cancer prognosis: a cohort study. *Cancer Cell International*, 18(1):1–10, 2018.

[170] Brian A Walker, Christopher P Wardell, Alex Murison, Eileen M Boyle, Dil B Begum, Nasrin M Dahir, Paula Z Proszek, Lorenzo Melchor, Charlotte Pawlyn, Martin F Kaiser, et al. Apobec family mutational signatures are associated with poor prognosis translocations in multiple myeloma. *Nature communications*, 6 (1):1–11, 2015.

[171] Phuc H Hoang, Alex J Cornish, Sara E Dobbins, Martin Kaiser, and Richard S Houlston. Mutational processes contributing to the development of multiple myeloma. *Blood cancer journal*, 9(8):1–11, 2019.

[172] Ayşen Timurağaoğlu, Sema Demircin, Seray Dizlek, Guchan Alanoğlu, and Evren Kiriş. Microsatellite instability is a common finding in multiple myeloma. *Clinical Lymphoma and Myeloma*, 9(5):371–374, 2009.

[173] Kaname Miyashita, Kei Fujii, Youko Suehiro, Kenichi Taguchi, Naokuni Uike, Mitsuaki A Yoshida, and Shinya Oda. Heterochronous occurrence of microsatellite instability in multiple myeloma–an implication for a role of defective dna mismatch repair in myelomagenesis. *Leukemia & Lymphoma*, 59(10):2454–2459, 2018.

[174] Liisa Chang, Minna Chang, Hanna M Chang, and Fuju Chang. Microsatellite instability: a predictive biomarker for cancer immunotherapy. *Applied Immunohistochemistry & Molecular Morphology*, 26(2):e15–e21, 2018.

[175] David Y Oh, Alan P Venook, and Lawrence Fong. On the verge: immunotherapy for colorectal carcinoma. *Journal of the National Comprehensive Cancer Network*, 13(8):970–978, 2015.

[176] Zhu Zhang, Bin Zhou, Qianqian Gao, Yuke Wu, Kui Zhang, Yan Pu, Yaping Song, Lin Zhang, and Mingrong Xi. A polymorphism at mirna-122-binding site in the il-1α 3 utr is associated with risk of epithelial ovarian cancer. *Familial cancer*, 13(4):595–601, 2014.

[177] X Chen, T Paranjape, C Stahlhut, T McVeigh, F Keane, S Nallur, N Miller, M Kerin, Y Deng, X Yao, et al. Targeted resequencing of the micrornaome and 3 utrome reveals functional germline dna variants with altered prevalence in epithelial ovarian cancer. *Oncogene*, 34(16):2125–2137, 2015.

[178] Ombretta Melaiu, Angelica Macauda, Juan Sainz, Diego Calvetti, Maria Sole Facioni, Giuseppe Maccari, Rob Ter Horst, Mihai G Netea, Yang Li, Norbert Grząśko, et al. Common gene variants within 3-untranslated regions as modulators of multiple myeloma risk and survival. *International Journal of Cancer*, 148(8):1887–1894, 2021.

[179] D Gareth R Evans, Elke M van Veen, Helen J Byers, Andrew J Wallace, Jamie M Ellingford, Glenda Beaman, Javier Santoyo-Lopez, Timothy J Aitman, Diana M Eccles, Fiona I Lalloo, et al. A dominantly inherited 5 utr variant causing methylation-associated silencing of brca1 as a cause of breast and ovarian cancer. *The American Journal of Human Genetics*, 103(2):213–220, 2018.

[180] Kaan Ozturk, Meltem Selen Onal, Ozgur Efiloglu, Emrah Nikerel, Asif Yildirim, and Dilek Telci. Association of 5 utr polymorphism of secretory phospholipase a2 group iia (pla2g2a) gene with prostate cancer metastasis. *Gene*, 742:144589, 2020.

[181] Brian Halbert and David J Einstein. Hot or not: tumor mutational burden (tmb) as a biomarker of immunotherapy response in genitourinary cancers. *Urology*, 147:119–126, 2021.

[182] Yanis Boumber. Tumor mutational burden (tmb) as a biomarker of response to immunotherapy in small cell lung cancer. *Journal of thoracic disease*, 10(8): 4689, 2018.

[183] Amber Miller, Laura Cattaneo, Yan W Asmann, Esteban Braggio, Jonathan J Keats, Daniel Auclair, Sagar Lonial, The MMRF CoMMpass Network, Stephen J Russell, and A Keith Stewart. Correlation between somatic mutation burden, neoantigen load and progression free survival in multiple myeloma: Analysis of mmrf commpass study. *Blood*, 128(22):193, 2016.

[184] JP Laubach, CE Paba Prada, PG Richardson, and DL Longo. Daratumumab, elotuzumab, and the development of therapeutic monoclonal antibodies in multiple myeloma. *Clinical Pharmacology & Therapeutics*, 101(1):81–88, 2017.

[185] Laura Moreno, Cristina Perez, Aintzane Zabaleta, Irene Manrique, Diego Alignani, Daniel Ajona, Laura Blanco, Marta Lasa, Patricia Maiso, Idoia Rodriguez, et al. The mechanism of action of the anti-cd38 monoclonal antibody isatuximab in multiple myeloma. *Clinical Cancer Research*, 25(10):3176–3187, 2019.

[186] Massimo Offidani, Laura Corvatta, Sonia Morè, and Attilio Olivieri. Belantamab mafodotin for the treatment of multiple myeloma: An overview of the clinical efficacy and safety. *Drug Design, Development and Therapy*, 15:2401, 2021.

[187] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[188] A. Farswan and A. Gupta. TV-DCT: Method to impute gene expression data using DCT based sparsity and total variation denoising. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1244–1248, May 2019. doi: 10.1109/ICASSP.2019.8683584.

[189] Shiv Gehlot, Akanksha Farswan, Anubha Gupta, and Ritu Gupta. CT-NNBI: Method to impute gene expression data using DCT based sparsity and Nuclear Norm constraint with Split Bregman Iteration. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1315–1318. IEEE, 2019.

[190] Akanksha Farswan, Anubha Gupta, Ritu Gupta, and Gurvinder Kaur. Imputation of gene expression data in blood cancer and its significance in inferring biological pathways. *Frontiers in oncology*, 9:1442, 2020.

[191] Jane R Montealegre, Renke Zhou, E Susan Amirian, and Michael E Scheurer. Uncovering nativity disparities in cancer patterns: Multiple imputation strategy to handle missing nativity data in the surveillance, epidemiology, and end results data file. *Cancer*, 120(8):1203–1211, 2014.

[192] Ritu Gupta, Gurvinder Kaur, Lalit Kumar, Lata Rani, Nitin Mathur, Atul Sharma, Meetu Dahiya, Varun Shekhar, Sadaf Khan, Anjali Mookerjee, et al. Nucleic acid based risk assessment and staging for clinical practice in multiple myeloma. *Annals of hematology*, 97(12):2447–2454, 2018.

[193] Shaji Kumar, Bruno Paiva, Kenneth C Anderson, Brian Durie, Ola Landgren, Philippe Moreau, Nikhil Munshi, Sagar Lonial, Joan Bladé, Maria-Victoria Mateos, et al. International myeloma working group consensus criteria for response and minimal residual disease assessment in multiple myeloma. *The lancet oncology*, 17(8):e328–e346, 2016.

[194] Christopher M Florkowski and Janice SC Chew-Harris. Methods of estimating gfr–different equations including ckd-epi. *The Clinical Biochemist Reviews*, 32 (2):75, 2011.

[195] S Vincent Rajkumar. Multiple myeloma: 2016 update on diagnosis, risk-stratification, and management. *American journal of hematology*, 91(7):719–734, 2016.

[196] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1):91–118, 2003.

[197] Gurvinder Kaur, Ritu Gupta, Nitin Mathur, Lata Rani, Lalit Kumar, Atul Sharma, Vishwajeet Singh, Anubha Gupta, and Om Dutt Sharma. Clinical impact of chromothriptic complex chromosomal rearrangements in newly diagnosed multiple myeloma. *Leukemia Research*, 76:58–64, 2019.

[198] Hervé Avet-Loiseau, Xavier Leleu, Murielle Roussel, Philippe Moreau, Catherine Guerin-Charbonnel, Denis Caillot, Gérald Marit, Lotfi Benboubker, Laurent Voillat, Claire Mathiot, et al. Bortezomib plus dexamethasone induction improves outcome of patients with t (4; 14) myeloma but not outcome of patients with del (17p). *Journal of Clinical Oncology*, 28(30):4630–4634, 2010.

[199] Akanksha Farswan, Anubha Gupta, Ritu Gupta, Saswati Hazra, Sadaf Khan, Lalit Kumar, and Atul Sharma. Ai-supported modified risk staging for multiple myeloma cancer useful in real-world scenario. *Translational oncology*, 14 (9):101157, 2021.

[200] Athira Unnikrishnan, Abdullah Mohammad Khan, Preeti Narayan, and Maxim Norkin. Striking age differences of multiple myeloma (mm) diagnosis in patients of indian and pakistani descent in the united states compared to native countries., 2017.

[201] AM Konatam and G Sadashivudu. Age of onset of multiple myeloma: a paradigm shift in indian patients. *Indian J Appl Res*, 6(3), 2016.

[202] Andriy Marusyk and Kornelia Polyak. Tumor heterogeneity: causes and consequences. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1805(1): 105–117, 2010.

[203] Maria Kleppe and Ross L Levine. Tumor heterogeneity confounds and illuminates: assessing the implications. *Nature medicine*, 20(4):342–344, 2014.

[204] Lauren MF Merlo, John W Pepper, Brian J Reid, and Carlo C Maley. Cancer as an evolutionary and ecological process. *Nature reviews cancer*, 6(12):924–935, 2006.

[205] Rebecca A Burrell, Nicholas McGranahan, Jiri Bartek, and Charles Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345, 2013.

[206] Li Ding, Timothy J Ley, David E Larson, Christopher A Miller, Daniel C Koboldt, John S Welch, Julie K Ritchey, Margaret A Young, Tamara Lamprecht, Michael D McLellan, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510, 2012.

[207] Christopher A Miller, Brian S White, Nathan D Dees, Malachi Griffith, John S Welch, Obi L Griffith, Ravi Vij, Michael H Tomasson, Timothy A Graubert, Matthew J Walter, et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS computational biology*, 10(8):e1003665, 2014.

[208] Mohammed El-Kebir, Layla Oesper, Hannah Acheson-Field, and Benjamin J Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70, 2015.

[209] Mohammed El-Kebir, Gryte Satas, and Benjamin J Raphael. Inferring parsimonious migration histories for metastatic cancers. *Nature genetics*, 50(5):718–726, 2018.

[210] Nicholas E Navin. Cancer genomics: one cell at a time. *Genome biology*, 15(8): 1–13, 2014.

[211] Devon A Lawson, Kai Kessenbrock, Ryan T Davis, Nicholas Pervolarakis, and Zena Werb. Tumour heterogeneity and metastasis at single-cell resolution. *Nature cell biology*, 20(12):1349–1360, 2018.

[212] Xun Xu, Yong Hou, Xuyang Yin, Li Bao, Aifa Tang, Luting Song, Fuqiang Li, Shirley Tsang, Kui Wu, Hanjie Wu, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, 148(5):886–895, 2012.

[213] Chang Yu, Jun Yu, Xiaotian Yao, William KK Wu, Youyong Lu, Senwei Tang, Xiangchun Li, Li Bao, Xiaoxing Li, Yong Hou, et al. Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing. *Cell research*, 24(6):701–712, 2014.

[214] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, 2011.

[215] Yong Wang, Jill Waters, Marco L Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513): 155–160, 2014.

[216] Yong Hou, Luting Song, Ping Zhu, Bo Zhang, Ye Tao, Xun Xu, Fuqiang Li, Kui Wu, Jie Liang, Di Shao, et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell*, 148(5):873–885, 2012.

[217] Hamim Zafar, Anthony Tzen, Nicholas Navin, Ken Chen, and Luay Nakhleh. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome biology*, 18(1):1–20, 2017.

[218] Charles Gawad, Winston Koh, and Stephen R Quake. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proceedings of the National Academy of Sciences*, 111(50):17947–17952, 2014.

[219] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome biology*, 17(1):1–17, 2016.

[220] Edith M Ross and Florian Markowetz. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome biology*, 17(1):1–14, 2016.

[221] Andrew Roth, Andrew McPherson, Emma Laks, Justina Biele, Damian Yap, Adrian Wan, Maia A Smith, Cydney B Nielsen, Jessica N McAlpine, Samuel Aparicio, et al. Clonal genotype and population structure inference from single-cell tumor sequencing. *Nature methods*, 13(7):573–576, 2016.

[222] Jochen Singer, Jack Kuipers, Katharina Jahn, and Niko Beerenwinkel. Single-cell mutation identification via phylogenetic inference. *Nature communications*, 9(1):1–8, 2018.

[223] Alexander Davis and Nicholas E Navin. Computing tumor trees from single cells. *Genome biology*, 17(1):1–4, 2016.

[224] Sayaka Miura, Louise A Huuki, Tiffany Buturla, Tracy Vu, Karen Gomez, and Sudhir Kumar. Computational enhancement of single-cell sequences for inferring tumor evolution. *Bioinformatics*, 34(17):i917–i926, 2018.

[225] Ziwei Chen, Fuzhou Gong, Lin Wan, and Liang Ma. RobustClone: a robust PCA method for tumor clone and evolution inference from single-cell sequencing data. *Bioinformatics*, 36(11):3299–3306, 2020.

[226] 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.

[227] Zhao Kang, Haiqi Pan, Steven CH Hoi, and Zenglin Xu. Robust graph learning from noisy data. *IEEE transactions on cybernetics*, 50(5):1833–1843, 2019.

[228] Nico Borgsmüller, Jose Bonet, Francesco Marass, Abel Gonzalez-Perez, Nuria Lopez-Bigas, and Niko Beerenwinkel. BnpC: Bayesian non-parametric clustering of single-cell mutation profiles. *Bioinformatics*, 36(19):4854–4859, 2020.

[229] Zhenhua Yu, Huidong Liu, Fang Du, and Xiaofen Tang. GRMT: generative reconstruction of mutation tree from scratch using single-cell sequencing data. *Frontiers in genetics*, page 970, 2021.

[230] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[231] Ludo Waltman and Nees Jan Van Eck. A smart local moving algorithm for large-scale modularity-based community detection. *The European physical journal B*, 86(11):1–14, 2013.

[232] Naoto Ozaki, Hiroshi Tezuka, and Mary Inaba. A simple acceleration method for the Louvain algorithm. *International Journal of Computer and Electrical Engineering*, 8(3):207, 2016.

[233] Seung-Hee Bae, Daniel Halperin, Jevin D West, Martin Rosvall, and Bill Howe. Scalable and efficient flow-based community detection for large-scale graph analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 11 (3):1–30, 2017.

[234] Vincent A Traag. Faster unfolding of communities: Speeding up the Louvain algorithm. *Physical Review E*, 92(3):032801, 2015.

[235] Yingrui Li, Xun Xu, Luting Song, Yong Hou, Zesong Li, Shirley Tsang, Fuqiang Li, Kate McGee Im, Kui Wu, Hanjie Wu, et al. Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. *GigaScience*, 1(1):2047–217X, 2012.

[236] Andrew McPherson, Andrew Roth, Emma Laks, Tehmina Masud, Ali Bashashati, Allen W Zhang, Gavin Ha, Justina Biele, Damian Yap, Adrian Wan, et al. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature genetics*, 48(7):758–767, 2016.

[237] Qiaojun Hong, Rong Li, Yiyan Zhang, and Kangsheng Gu. Fibrillin 2 gene knockdown inhibits invasion and migration of lung cancer cells. *Cellular and Molecular Biology*, 66(7):190–196, 2020.

Appendix A

Missing value imputation in gene expression data

This Section contains additional results of Chapter 1 on imputation of gene expression matrices in the form of tables and bargraphs on the four dataset- CLL, AML, MM-Spanish and MM-India. It contains results obtained from KEGG pathway analysis. These tables and figures are appropriately referred in the main manuscript.

Table A.1: Classification accuracy and F1 scores on different sampling percentage of incomplete matrix and the recovered/imputed matrix on MM-Spanish data. SR stands for Sampling ratio of observed data to the total data (in percentage)

	Classification Accuracy				**F1 score**			
	KNN		**Linear SVM classifier**		**KNN**		**Linear SVM classifier**	
SR	Observed	Recovered	Observed	Recovered	Observed	Recovered	Observed	Recovered
10	0.22	0.6	0.77	0.8	0.21	0.59	0.75	0.8
20	0.35	0.82	0.7	0.88	0.37	0.82	0.68	0.88
30	0.6	0.88	0.72	0.9	0.6	0.88	0.71	0.9
40	0.56	0.86	0.65	0.92	0.54	0.84	0.64	0.92
50	0.54	0.84	0.72	0.94	0.65	0.84	0.7	0.94
60	0.66	0.83	0.76	0.93	0.63	0.84	0.74	0.93
70	0.63	0.89	0.75	0.92	0.62	0.89	0.74	0.92
80	0.67	0.86	0.83	0.92	0.86	0.86	0.81	0.92
90	0.67	0.87	0.86	0.94	0.88	0.87	0.85	0.94

Table A.2: Adjusted p-values for KEGG pathways at ground truth, 50% observed and imputed data, 70% observed and imputed data for CLL dataset.

		50%		**70%**	
Term	**Ground Truth**	**Observed**	**Recovered**	**Observed**	**Recovered**
Pathways in cancer	2.75e-13	6.17e-07	1.08e-08	1.21e-11	5.59e-13
Transcriptional misregulation in cancer	3.30e-15	3.25e-15	1.82e-12	3.26e-15	6.98e-12
MAPK signaling pathway	3.13e-08	5.71e-06	1.50e-05	5.13e-12	5.25e-10
Acute myeloid leukemia	3.04e-08	3.31e-03	1.57e-04	3.01e-08	2.79e-07
T cell receptor signaling pathway	2.55e-08	2.28e-01	2.25e-01	2.40e-04	1.48e-05
Chronic myeloid leukemia	5.54e-08	9.35e-03	3.88e-05	3.44e-07	7.84e-09
HTLV-I infection	1.42e-07	5.35e-05	5.22e-07	5.13e-12	6.34e-07
Proteoglycans in cancer	5.69e-07	1.51e-03	1.85e-03	2.21e-05	8.40e-06
ErbB signaling pathway	2.56e-06	1.51e-01	2.73e-03	5.51e-05	5.09e-05
B cell receptor signaling pathway	1.66e-05	2.14e-01	2.11e-01	1.49e-03	1.40e-03
AGE-RAGE signaling pathway	5.93e-05	4.02e-02	1.70e-02	1.91e-04	4.35e-05
Neurotrophin signaling pathway	6.74e-05	4.82e-03	3.88e-05	1.27e-07	3.38e-06
Wnt signaling pathway	8.33e-05	3.10e-02	2.13e-03	6.66e-05	1.90e-05
Signaling pathways regulating pluripotency of stem cells	8.33e-05	6.01e-04	3.74e-06	5.33e-06	6.18e-05
Chemokine signaling pathway	1.13e-04	2.02e-01	2.47e-02	3.82e-02	8.19e-02
TNF signaling pathway	1.13e-04	2.81e-04	2.25e-02	1.94e-07	1.27e-06
TGF-beta signaling pathway	2.37e-04	1.27e-04	4.75e-04	4.27e-05	1.81e-04
Toll-like receptor signaling pathway	3.17e-04	6.52e-03	1.14e-01	1.57e-05	3.99e-06
PI3K-Akt signaling pathway	1.43e-03	2.71e-03	2.14e-02	3.31e-05	1.99e-04
p53 signaling pathway	2.67e-04	2.23e-03	2.47e-02	2.22e-04	9.81e-04
HIF-1 signaling pathway	1.02e-03	6.00e-03	1.82e-02	1.24e-05	1.99e-04
NF-kappa B signaling pathway	4.65e-04	1.48e-05	3.53e-03	9.93e-05	2.19e-05
Ras signaling pathway	1.96e-03	3.82e-02	3.99e-03	7.12e-04	4.73e-03
Jak-STAT signaling pathway	1.97e-03	2.03e-01	2.02e-01	1.49e-02	5.35e-03
Hippo signaling pathway	4.57e-03	1.10e-03	3.99e-04	1.49e-03	4.28e-04
FoxO signaling pathway	5.12e-03	5.27e-02	2.25e-02	1.56e-03	3.59e-05
Phospholipase D signaling pathway	8.76e-03	6.70e-01	1.50e-01	1.16e-01	5.33e-02
RNA degradation	2.04e-03	3.99e-03	1.45e-02	1.96e-03	8.58e-07
cGMP-PKG signaling pathway	2.17e-02	7.95e-01	7.87e-01	4.66e-02	4.79e-02
Insulin signaling pathway	1.96e-02	6.49e-01	6.39e-02	1.02e-01	1.81e-02
cAMP signaling pathway	5.82e-02	5.60e-01	2.36e-01	1.05e-02	5.33e-02
NOD-like receptor signaling pathway	1.74e-03	3.31e-03	1.21e-01	1.56e-03	1.18e-05
Fc epsilon RI signaling pathway	1.72e-02	3.98e-01	1.83e-01	5.07e-02	1.60e-02
RIG-I-like receptor signaling pathway	6.08e-02	2.02e-01	7.16e-01	1.79e-02	4.73e-03
Notch signaling pathway	2.00e-01	2.95e-04	6.94e-03	5.76e-02	1.56e-02

Table A.3: Adjusted p-values for KEGG pathways at ground truth, 50% observed and imputed data, 70% observed and imputed data for AML dataset.

		50%		**70%**	
Term	**Ground Truth**	**Observed**	**Recovered**	**Observed**	**Recovered**
Signaling pathways regulating pluripotency of stem cells	2.76e-12	2.06e-19	1.63e-17	1.57e-15	1.02e-12
Pathways in cancer	2.71e-09	6.56e-10	8.63e-06	2.81e-11	2.81e-11
HTLV-I infection	1.58e-05	5.59e-07	2.86e-04	2.34e-06	2.34e-06
Hippo signaling pathway	9.62e-06	2.21e-06	3.11e-05	2.99e-07	7.10e-06
Transcriptional misregulation in cancer	1.58e-05	5.42e-04	1.37e-07	2.71e-08	2.71e-08
TGF-beta signaling pathway	2.47e-05	9.89e-07	1.15e-06	2.30e-09	1.15e-06
FoxO signaling pathway	1.21e-04	6.34e-06	4.98e-02	6.27e-06	1.23e-06
Proteoglycans in cancer	6.57e-04	7.41e-05	1.89e-04	2.00e-07	2.00e-07
Hedgehog signaling pathway	4.57e-04	7.41e-05	1.48e-03	2.31e-04	8.04e-03
Cytokine-cytokine receptor interaction	1.33e-03	1.27e-02	7.11e-03	1.09e-03	3.57e-06
p53 signaling pathway	6.57e-04	5.56e-04	1.66e-03	1.66e-03	5.31e-05
PI3K-Akt signaling pathway	2.97e-03	2.53e-04	1.02e-03	1.02e-03	5.26e-07
Wnt signaling pathway	2.02e-03	5.88e-04	8.98e-04	2.65e-02	2.40e-04
T cell receptor signaling pathway	2.24e-03	2.36e-03	4.64e-02	4.58e-03	1.19e-03
Jak-STAT signaling pathway	4.38e-03	1.61e-04	6.49e-03	6.36e-04	6.36e-04
HIF-1 signaling pathway	6.72e-03	1.61e-04	5.36e-05	4.32e-03	1.02e-05
ErbB signaling pathway	8.50e-03	2.30e-02	6.12e-03	6.12e-03	1.50e-03
Neurotrophin signaling pathway	1.67e-02	8.99e-02	8.08e-02	1.06e-02	1.06e-02
MAPK signaling pathway	2.47e-02	3.26e-04	1.23e-02	2.49e-04	8.05e-05
Ras signaling pathway	4.98e-02	1.89e-02	2.09e-01	6.22e-04	4.70e-06
Toll-like receptor signaling pathway	6.60e-02	1.33e-01	2.74e-01	1.27e-01	1.71e-02
AGE-RAGE signaling pathway	1.20e-01	5.35e-03	7.22e-01	3.83e-03	4.47e-05
cGMP-PKG signaling pathway	1.60e-01	4.41e-01	7.91e-01	2.40e-01	2.48e-02
TNF signaling pathway	1.60e-01	1.48e-01	7.65e-01	2.05e-02	6.40e-03
Rap1 signaling pathway	2.00e-01	1.99e-01	4.33e-01	7.02e-03	2.42e-05
NOD-like receptor signaling pathway	5.43e-02	1.40e-01	4.19e-01	5.40e-02	2.92e-03
AMPK signaling pathway	2.12e-01	5.91e-03	3.98e-03	1.27e-02	2.85e-04
Fc epsilon RI signaling pathway	2.12e-01	2.54e-02	5.10e-01	9.04e-02	2.75e-02

Table A.4: Adjusted *p*-values for KEGG pathways at ground truth, 50% observed and imputed data, 70% observed and imputed data for MM-Spanish dataset.

Term		**50%**		**70%**	
	Ground Truth	**Observed**	**Recovered**	**Observed**	**Recovered**
Epstein-Barr virus infection	6.07e-08	8.04e-03	2.16e-03	3.61e-04	8.59e-07
T cell receptor signaling pathway	2.31e-05	6.10e-02	2.76e-02	2.34e-03	1.41e-03
Pathways in cancer	1.18e-04	1.92e-04	5.94e-09	3.37e-03	2.07e-05
RNA polymerase	4.92e-05	3.32e-01	3.43e-01	3.47e-01	3.84e-02
RNA degradation	1.18e-04	5.45e-02	4.26e-04	4.60e-02	7.65e-03
MAPK signaling pathway	4.50e-04	1.40e-01	2.81e-02	2.58e-02	4.67e-02
NF-kappa B signaling pathway	4.10e-04	5.83e-01	9.61e-02	3.57e-02	4.04e-02
FoxO signaling pathway	1.59e-03	6.95e-02	1.82e-03	1.54e-03	2.56e-03
Apoptosis	2.22e-03	7.94e-01	3.99e-02	7.59e-02	1.67e-02
Neurotrophin signaling pathway	2.22e-03	9.39e-02	2.25e-03	4.08e-02	4.51e-02
HTLV-I infection	2.62e-03	9.13e-03	2.36e-01	8.16e-03	1.18e-03
Ras signaling pathway	5.26e-03	1.46e-01	2.81e-02	2.37e-01	4.67e-02
Rap1 signaling pathway	6.37e-03	3.65e-02	3.36e-02	1.83e-01	5.83e-02
Thyroid hormone signaling pathway	6.22e-03	2.80e-03	2.16e-03	5.64e-04	1.25e-03
PI3K-Akt signaling pathway	6.62e-03	1.91e-02	4.02e-03	3.37e-03	1.58e-02
AGE-RAGE signaling pathway in diabetic complications	6.59e-03	2.65e-02	2.38e-02	1.07e-01	9.55e-03
VEGF signaling pathway	6.25e-03	1.58e-01	6.69e-02	1.55e-01	1.53e-01
RNA transport	7.73e-03	2.74e-01	1.66e-01	8.34e-02	1.07e-02
HIF-1 signaling pathway	6.89e-03	1.20e-01	9.45e-03	2.50e-02	2.31e-01
Transcriptional misregulation in cancer	1.03e-02	8.04e-03	2.91e-04	5.41e-02	1.33e-03
Fc epsilon RI signaling pathway	8.37e-03	6.67e-01	9.37e-02	1.94e-01	1.94e-01
Signaling pathways regulating pluripotency of stem cells	1.43e-02	5.00e-03	8.24e-03	3.57e-03	9.28e-03
Prolactin signaling pathway	1.09e-02	4.33e-01	4.07e-02	3.68e-02	4.16e-02
ErbB signaling pathway	2.54e-02	1.58e-01	3.08e-02	3.96e-03	1.12e-02
cAMP signaling pathway	4.16e-02	5.00e-03	3.50e-05	2.50e-02	4.67e-02
B cell receptor signaling pathway	3.66e-02	2.26e-01	4.23e-02	9.69e-02	1.04e-01

Table A.5: Adjusted *p*-values for KEGG pathways at ground truth and 70% observed and imputed data for MM-Indian dataset.

Term		**70%**	
	Adjusted *p*-value	**Observed**	**Recovered**
Signaling pathways regulating pluripotency of stem cells	3.23e-03	3.31e-04	2.26e-05
Pathways in cancer	4.65e-03	3.43e-04	1.00e-04
PI3K-Akt signaling pathway	3.30e-03	5.46e-03	7.06e-04
Proteoglycans in cancer	3.72e-03	5.68e-04	7.84e-05
Ras signaling pathway	4.93e-03	2.42e-02	1.54e-02
Breast cancer	4.33e-03	4.45e-04	1.53e-03
Gastric cancer	4.26e-03	6.27e-05	8.14e-05
T cell receptor signaling pathway	8.44e-03	8.61e-02	3.07e-03
Non-small cell lung cancer	7.95e-03	5.50e-03	1.12e-02
Transcriptional misregulation in cancer	7.82e-03	5.55e-01	1.37e-03
ErbB signaling pathway	7.59e-03	5.57e-03	3.48e-03
Regulation of actin cytoskeleton	8.59e-03	1.05e-01	1.36e-01
Spliceosome	1.31e-02	3.52e-04	1.51e-02
Pancreatic cancer	1.23e-02	2.89e-03	1.76e-02
Herpes simplex virus 1 infection	1.41e-02	5.43e-03	5.48e-04
Human papillomavirus infection	1.33e-02	1.02e-02	1.53e-03
HIF-1 signaling pathway	1.63e-02	3.36e-02	2.96e-05
Hepatocellular carcinoma	1.86e-02	3.17e-04	5.62e-04
Acute myeloid leukemia	2.07e-02	1.64e-02	7.61e-04
Colorectal cancer	2.14e-02	2.17e-03	3.30e-04
Renal cell carcinoma	2.43e-02	1.96e-02	3.48e-04
TGF-beta signaling pathway	2.49e-02	5.52e-02	9.17e-02
Rap1 signaling pathway	2.71e-02	4.73e-01	5.95e-02
Melanoma	2.62e-02	3.84e-01	1.52e-02
Primary immunodeficiency	2.55e-02	6.56e-01	5.53e-02
Cellular senescence	2.67e-02	3.84e-01	7.30e-03
Thyroid hormone signaling pathway	2.80e-02	1.36e-01	4.13e-02
Chronic myeloid leukemia	3.07e-02	9.44e-03	1.82e-02
Prostate cancer	3.06e-02	4.02e-03	7.19e-03
Sphingolipid signaling pathway	3.01e-02	6.91e-02	3.43e-04
Cell cycle	3.83e-02	1.67e-01	1.21e-01

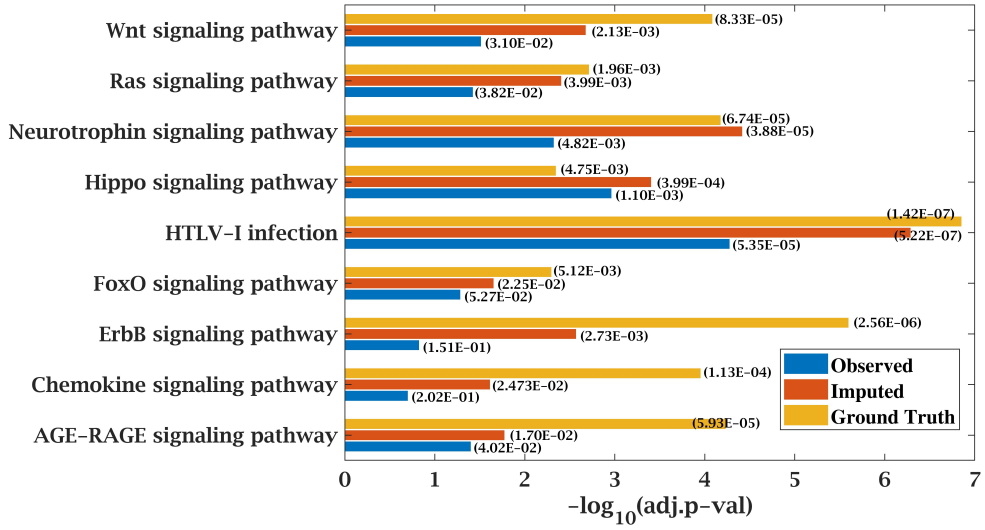


Figure A.1: Few important KEGG pathways at 50% observed and imputed data for CLL data. Adjusted p-values are shown in brackets.

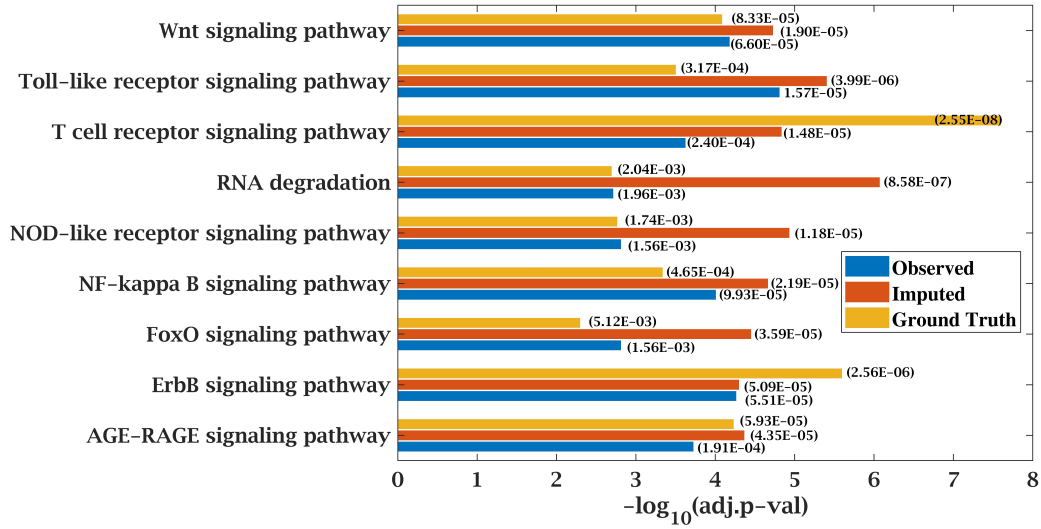


Figure A.2: Few important KEGG pathways at 70% observed and imputed data for CLL data. Adjusted p-values are shown in brackets.

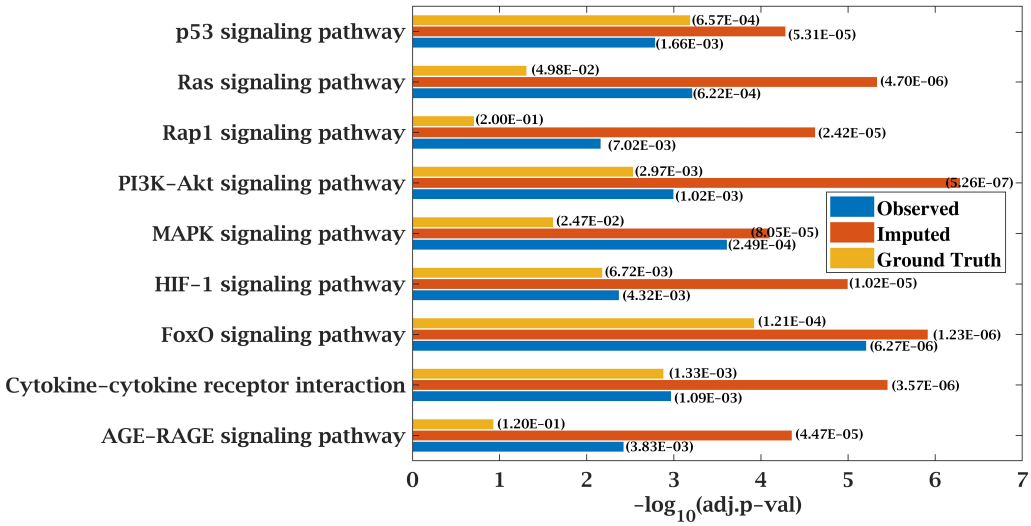


Figure A.3: Few important KEGG pathways at 70% observed and imputed data for AML data. Adjusted p-values are shown in brackets

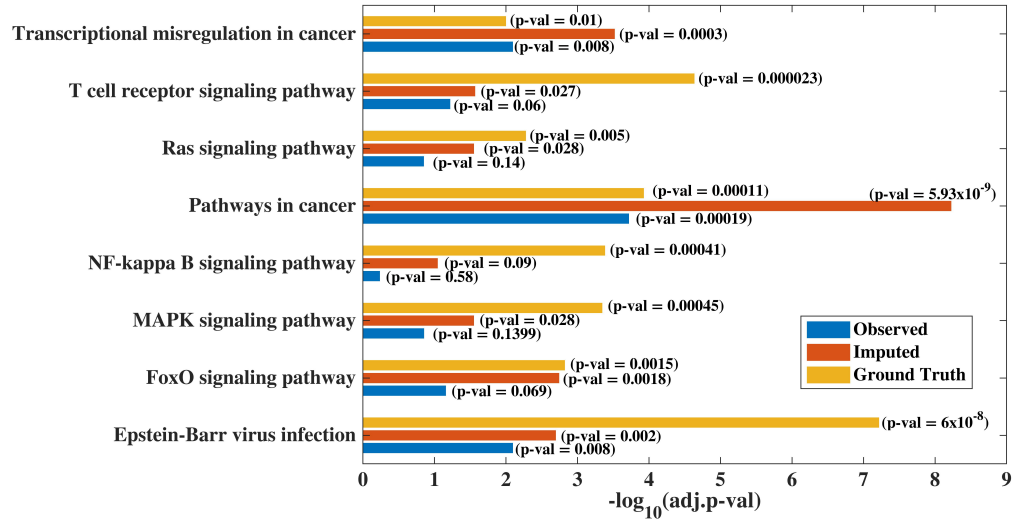


Figure A.4: Few important KEGG pathways at 50% observed and imputed data for MM- Spanish data. Adjusted p-values are shown in brackets.

Table A.6: KEGG pathways on CLL dataset

	100%	50%		70%	
	Ground Truth	Observed	Imputed	Observed	Imputed
1	**Transcriptional misregulation in cancer_Homo sapiens_hsa05202**				
Overlap	31/180	31/180	28/180	31/180	27/180
Adj. p-value	3.30e-15	3.25e-15	1.82e-12	3.26e-15	6.98e-12
Comb. score	67.49	67.49	56.44	67.49	52.19
2	**Pathways in cancer_Homo sapiens_hsa05200**				
Overlap	42/397	32/397	35/397	39/397	42/397
Adj. p-value	2.75e-13	6.17e-07	1.08e-08	1.21e-11	5.59e-13
Comb. score	69.69	39.19	47.65	57.57	70.58
3	**Hepatitis B_Homo sapiens_hsa05161**				
Overlap	24/146	14/146	15/146	24/146	24/146
Adj. p-value	1.45e-11	2.80e-04	6.18e-05	1.08e-11	1.47e-11
Comb. score	54.83	18.06	20.59	54.09	54.83
4	**Measles_Homo sapiens_hsa05162**				
Overlap	23/136	10/136	8/136	15/136	16/136
Adj. p-value	1.86e-11	9.35e-03	5.86e-02	1.20e-05	3.38e-06
Comb. score	50.89	4.87	0.75	16.69	22.20
5	**Herpes simplex infection_Homo sapiens_hsa05168**				
Overlap	23/185	14/185	13/185	21/185	23/185
Adj. p-value	9.76e-09	1.83e-03	4.88e-03	1.27e-07	7.84e-09
Comb. score	37.11	10.24	7.75	28.19	36.58
6	**T cell receptor signaling pathway_Homo sapiens_hsa04660**				
Overlap	17/104	5/104	5/104	11/104	13/104
Adj. p-value	2.55e-08	2.28e-01	2.25e-01	2.40e-04	1.48e-05
Comb. score	36.46	-1.59	-1.76	6.47	15.99
7	**Acute myeloid leukemia_Homo sapiens_hsa05221**				
Overlap	13/57	7/57	9/57	13/57	12/57
Adj. p-value	3.04e-08	3.31e-03	1.57e-04	3.01e-08	2.79e-07
Comb. score	36.94	7.01	17.66	36.94	31.04
8	**MAPK signaling pathway_Homo sapiens_hsa04010**				
Overlap	26/255	23/255	22/255	32/255	29/255
Adj. p-value	3.13e-08	5.71e-06	1.50e-05	5.13e-12	5.25e-10
Comb. score	37.06	30.54	27.12	60.04	48.47
9	**Epstein-Barr virus infection_Homo sapiens_hsa05169**				
Overlap	23/202	18/202	18/202	21/202	25/202
Adj. p-value	3.13e-08	7.06e-05	5.87e-05	4.25e-07	1.81e-09
Comb. score	35.55	22.01	21.38	26.48	43.94
10	**Chronic myeloid leukemia_Homo sapiens_hsa05220**				
Overlap	14/73	7/73	11/73	13/73	15/73
Adj. p-value	5.54e-08	9.35e-03	3.88e-05	3.44e-07	7.84e-09
Comb. score	32.25	3.22	22.10	25.72	38.24
11	**HTLV-I infection_Homo sapiens_hsa05166**				
Overlap	25/258	21/258	25/258	32/258	24/258
Adj. p-value	1.42e-07	5.35e-05	5.22e-07	5.13e-12	6.34e-07
Comb. score	31.48	22.93	35.20	57.00	28.98
12	**Viral carcinogenesis_Homo sapiens_hsa05203**				
Overlap	22/205	14/205	11/205	24/205	18/205
Adj. p-value	1.61e-07	3.99e-03	4.05e-02	1.15e-08	2.61e-05
Comb. score	31.30	8.90	1.39	40.27	15.07
13	**Thyroid cancer_Homo sapiens_hsa05216**				
Overlap	9/29	6/29	9/29	6/29	9/29
Adj. p-value	3.26e-07	6.64e-04	1.06e-06	2.59e-04	4.31e-07
Comb. score	19.55	7.12	29.27	-10.53	22.79
14	**Proteoglycans in cancer_Homo sapiens_hsa05205**				
Overlap	21/203	15/203	15/203	18/203	19/203
Adj. p-value	5.69e-07	1.51e-03	1.85e-03	2.21e-05	8.40e-06
Comb. score	28.45	12.38	13.27	15.46	21.20
15	**Inflammatory bowel disease (IBD)_Homo sapiens_hsa05321**				
Overlap	12/65	11/65	11/65	13/65	14/65
Adj. p-value	7.84e-07	1.48e-05	1.50e-05	1.27e-07	1.39e-08
Comb. score	23.71	25.06	25.06	32.13	36.19
16	**Pancreatic cancer_Homo sapiens_hsa05212**				
Overlap	12/66	7/66	9/66	13/66	12/66
Adj. p-value	8.78e-07	6.48e-03	4.31e-04	1.27e-07	1.10e-06
Comb. score	23.89	5.03	14.86	31.33	25.70

Table A.7: KEGG pathways on CLL dataset continued from Table A.6

	100%	50%		70%	
	Ground Truth	Observed	Imputed	Observed	Imputed
17	**ErbB signaling pathway_Homo sapiens_hsa04012**				
Overlap	13/87	5/87	9/87	11/87	11/87
Adj. p-value	2.56e-06	1.51e-01	2.73e-03	5.51e-05	5.09e-05
Comb. score	21.71	-1.29	10.26	9.84	8.48
18	**Spliceosome_Homo sapiens_hsa03040**				
Overlap	16/134	16/134	13/134	15/134	13/134
Adj. p-value	2.56e-06	9.06e-06	3.99e-04	1.08e-05	1.35e-04
Comb. score	20.95	25.83	15.02	16.95	7.37
19	**Osteoclast differentiation_Homo sapiens_hsa04380**				
Overlap	15/132	14/132	11/132	17/132	13/132
Adj. p-value	1.09e-05	1.05e-04	3.38e-03	3.88e-07	1.18e-04
Comb. score	19.17	19.68	9.28	25.39	7.89
20	**Hepatitis C_Homo sapiens_hsa05160**				
Overlap	15/133	8/133	9/133	15/133	14/133
Adj. p-value	1.14e-05	5.27e-02	2.25e-02	1.06e-05	3.59e-05
Comb. score	18.38	0.78	2.68	18.03	12.18
21	**B cell receptor signaling pathway_Homo sapiens_hsa04662**				
Overlap	11/73	4/73	4/73	8/73	8/73
Adj. p-value	1.66e-05	2.14e-01	2.11e-01	1.49e-03	1.40e-03
Comb. score	17.46	-2.50	-2.73	1.12	-0.69
22	**Influenza A_Homo sapiens_hsa05164**				
Overlap	17/175	10/175	9/175	16/175	16/175
Adj. p-value	1.69e-05	3.91e-02	7.88e-02	4.70e-05	4.44e-05
Comb. score	18.71	1.93	0.10	12.99	12.09
23	**Colorectal cancer_Homo sapiens_hsa05210**				
Overlap	10/62	5/62	6/62	9/62	10/62
Adj. p-value	2.36e-05	5.35e-02	1.84e-02	1.05e-04	1.90e-05
Comb. score	14.51	-2.20	1.47	5.44	11.49
24	**Renal cell carcinoma_Homo sapiens_hsa05211**				
Overlap	10/66	7/66	8/66	12/66	9/66
Adj. p-value	4.06e-05	6.48e-03	2.13e-03	7.72e-07	1.52e-04
Comb. score	13.32	4.44	10.37	21.49	3.14
25	**AGe-RAGE signaling pathway in diabetic complications_Homo sapiens_hsa04933**				
Overlap	12/101	7/101	8/101	11/101	12/101
Adj. p-value	5.93e-05	4.02e-02	1.70e-02	1.91e-04	4.35e-05
Comb. score	15.76	1.49	4.95	8.62	12.41
26	**Legionellosis_Homo sapiens_hsa05134**				
Overlap	9/55	11/55	7/55	10/55	9/55
Adj. p-value	5.93e-05	5.71e-06	3.26e-03	8.55e-06	4.35e-05
Comb. score	10.66	29.69	7.40	15.56	5.72
27	**Neurotrophin signaling pathway_Homo sapiens_hsa04722**				
Overlap	13/120	10/120	14/120	17/120	15/120
Adj. p-value	6.74e-05	4.82e-03	3.88e-05	1.27e-07	3.38e-06
Comb. score	13.78	7.30	22.43	30.22	21.43
28	**Wnt signaling pathway_Homo sapiens_hsa04310**				
Overlap	14/142	9/142	12/142	14/142	15/142
Adj. p-value	8.33e-05	3.10e-02	2.13e-03	6.66e-05	1.90e-05
Comb. score	13.11	2.25	11.43	10.83	15.26
29	**Signaling pathways regulating pluripotency of stem cells_hsa04550**				
Overlap	14/142	13/142	17/142	16/142	14/142
Adj. p-value	8.33e-05	6.01e-04	3.74e-06	5.33e-06	6.18e-05
Comb. score	12.96	13.68	27.86	19.89	9.32
30	**Prostate cancer_Homo sapiens_hsa05215**				
Overlap	11/89	5/89	7/89	12/89	13/89
Adj. p-value	8.33e-05	1.59e-01	2.37e-02	1.47e-05	3.38e-06
Comb. score	12.36	-1.36	1.84	15.57	22.43
31	**Insulin resistance_Homo sapiens_hsa04931**				
Overlap	12/109	9/109	9/109	13/109	12/109
Adj. p-value	1.09e-04	7.23e-03	8.58e-03	2.06e-05	7.99e-05
Comb. score	11.64	4.96	5.98	13.78	8.17
32	**Chemokine signaling pathway_Homo sapiens_hsa04062**				
Overlap	16/187	8/187	11/187	10/187	9/187
Adj. p-value	1.13e-04	2.02e-01	2.47e-02	3.82e-02	8.19e-02
Comb. score	11.88	-0.82	2.32	-1.91	-2.45

Table A.8: KEGG pathways on CLL dataset (continued from Table A.7

	100%	50%		70%	
	Ground Truth	Observed	Imputed	Observed	Imputed
33	**TNF signaling pathway_Homo sapiens_hsa04668**				
Overlap	12/110	12/110	8/110	16/110	15/110
Adj. p-value	1.13e-04	2.81e-04	2.25e-02	1.94e-07	1.27e-06
Comb. score	11.69	17.22	2.60	28.21	24.78
34	**Cytosolic DNA-sensing pathway_Homo sapiens_hsa04623**				
Overlap	9/64	6/64	6/64	7/64	7/64
Adj. p-value	1.58e-04	1.99e-02	2.02e-02	2.87e-03	3.04e-03
Comb. score	5.45	0.03	0.22	-5.01	-4.50
35	**Endometrial cancer_Homo sapiens_hsa05213**				
Overlap	8/52	5/52	8/52	6/52	7/52
Adj. p-value	2.19e-04	3.10e-02	4.52e-04	4.41e-03	9.32e-04
Comb. score	5.51	-1.45	13.31	-7.65	-4.28
36	**TGF-beta signaling pathway_Homo sapiens_hsa04350**				
Overlap	10/84	11/84	10/84	11/84	10/84
Adj. p-value	2.37e-04	1.27e-04	4.75e-04	4.27e-05	1.81e-04
Comb. score	8.13	18.51	13.58	10.66	4.50
37	**p53 signaling pathway_Homo sapiens_hsa04115**				
Overlap	9/69	8/69	6/69	9/69	8/69
Adj. p-value	2.67e-04	2.23e-03	2.47e-02	2.22e-04	9.81e-04
Comb. score	5.17	7.30	-0.65	2.78	-2.47
38	**Small cell lung cancer_Homo sapiens_hsa05222**				
Overlap	10/86	4/86	4/86	8/86	12/86
Adj. p-value	2.76e-04	2.94e-01	3.01e-01	3.51e-03	1.24e-05
Comb. score	7.16	-2.41	-2.36	-2.15	16.50
39	**MicroRNAs in cancer_Homo sapiens_hsa05206**				
Overlap	20/297	13/297	17/297	20/297	20/297
Adj. p-value	3.03e-04	8.87e-02	7.41e-03	2.54e-04	2.31e-04
Comb. score	9.57	0.32	7.32	7.94	6.54
40	**Cell cycle_Homo sapiens_hsa04110**				
Overlap	12/124	11/124	9/124	12/124	16/124
Adj. p-value	3.07e-04	2.08e-03	1.72e-02	2.56e-04	1.12e-06
Comb. score	7.19	9.20	3.71	5.49	22.90
41	**Toll-like receptor signaling pathway_Homo sapiens_hsa04620**				
Overlap	11/106	9/106	6/106	13/106	14/106
Adj. p-value	3.17e-04	6.52e-03	1.14e-01	1.57e-05	3.99e-06
Comb. score	7.80	5.91	-0.90	15.10	20.81
42	**Leishmaniasis_Homo sapiens_hsa05140**				
Overlap	9/73	7/73	6/73	8/73	9/73
Adj. p-value	3.69e-04	9.35e-03	3.05e-02	1.49e-03	2.92e-04
Comb. score	4.85	2.53	-0.37	0.14	1.11
43	**Pertussis_Homo sapiens_hsa05133**				
Overlap	9/75	9/75	6/75	11/75	10/75
Adj. p-value	4.46e-04	9.43e-04	3.35e-02	1.57e-05	7.58e-05
Comb. score	3.54	10.33	-1.02	10.79	4.59
44	**NF-kappa B signaling pathway_Homo sapiens_hsa04064**				
Overlap	10/93	13/93	9/93	11/93	12/93
Adj. p-value	4.65e-04	1.48e-05	3.53e-03	9.93e-05	2.19e-05
Comb. score	4.97	23.47	7.54	7.73	11.86
45	**Tuberculosis_Homo sapiens_hsa05152**				
Overlap	14/178	14/178	12/178	16/178	15/178
Adj. p-value	6.64e-04	1.34e-03	8.98e-03	5.51e-05	1.73e-04
Comb. score	6.40	11.29	5.87	11.16	6.87
46	**Toxoplasmosis_Homo sapiens_hsa05145**				
Overlap	11/118	9/118	8/118	14/118	14/118
Adj. p-value	7.44e-04	1.14e-02	3.03e-02	1.15e-05	1.24e-05
Comb. score	5.83	3.97	1.47	17.30	17.67
47	**HIF-1 signaling pathway_Homo sapiens_hsa04066**				
Overlap	10/103	9/103	8/103	13/103	11/103
Adj. p-value	1.02e-03	6.00e-03	1.82e-02	1.24e-05	1.99e-04
Comb. score	5.01	6.50	3.74	15.51	4.90
48	**Chagas disease (American trypanosomiasis)_Homo sapiens_hsa05142**				
Overlap	10/104	11/104	8/104	16/104	13/104
Adj. p-value	1.08e-03	6.01e-04	1.85e-02	1.27e-07	1.48e-05
Comb. score	4.54	13.62	3.28	30.48	15.87

Table A.9: KEGG pathways on CLL dataset (continued from Table A.8)

	100%	50%		70%	
	Ground Truth	Observed	Imputed	Observed	Imputed
49	**Epithelial cell signaling in Helicobacter pylori infection_Homo sapiens_hsa05120**				
Overlap	8/68	7/68	5/68	8/68	9/68
Adj. p-value	1.09e-03	6.79e-03	7.11e-02	9.75e-04	1.82e-04
Comb. score	1.45	3.08	-2.43	-0.06	1.77
50	**Salmonella infection_Homo sapiens_hsa05132**				
Overlap	9/86	8/86	7/86	9/86	11/86
Adj. p-value	1.10e-03	6.52e-03	2.14e-02	9.91e-04	4.67e-05
Comb. score	2.16	3.90	1.37	0.95	6.85
51	**Melanoma_Homo sapiens_hsa05218**				
Overlap	8/71	6/71	8/71	9/71	9/71
Adj. p-value	1.41e-03	3.02e-02	3.12e-03	2.56e-04	2.38e-04
Comb. score	2.02	1.06	9.51	3.45	1.81
52	**PI3K-Akt signaling pathway_Homo sapiens_hsa04151**				
Overlap	20/341	20/341	17/341	24/341	22/341
Adj. p-value	1.43e-03	2.71e-03	2.14e-02	3.31e-05	1.99e-04
Comb. score	5.83	10.90	4.25	15.24	7.85
53	**Adherens junction_Homo sapiens_hsa04520**				
Overlap	8/74	6/74	6/74	4/74	3/74
Adj. p-value	1.73e-03	3.36e-02	3.19e-02	1.81e-01	4.20e-01
Comb. score	0.42	-0.06	-0.67	-4.33	-2.99
54	**Longevity regulating pathway - mammal_Homo sapiens_hsa04211**				
Overlap	9/94	7/94	8/94	11/94	9/94
Adj. p-value	1.88e-03	3.10e-02	1.20e-02	1.05e-04	1.70e-03
Comb. score	2.20	1.79	5.09	8.44	0.68
55	**Natural killer cell mediated cytotoxicity_Homo sapiens_hsa04650**				
Overlap	11/135	3/135	5/135	6/135	7/135
Adj. p-value	1.89e-03	8.45e-01	3.93e-01	1.92e-01	9.30e-02
Comb. score	2.96	-0.72	-1.57	-2.36	-2.56
56	**Ras signaling pathway_Homo sapiens_hsa04014**				
Overlap	15/227	12/227	15/227	16/227	14/227
Adj. p-value	1.96e-03	3.82e-02	3.99e-03	7.12e-04	4.73e-03
Comb. score	3.75	2.36	9.48	5.87	1.42
57	**Jak-STAT signaling pathway_Homo sapiens_hsa04630**				
Overlap	12/158	7/158	7/158	10/158	11/158
Adj. p-value	1.97e-03	2.03e-01	2.02e-01	1.49e-02	5.35e-03
Comb. score	3.17	-0.82	-0.99	-0.39	0.30
58	**Thyroid hormone signaling pathway_Homo sapiens_hsa04919**				
Overlap	10/118	10/118	7/118	10/118	10/118
Adj. p-value	2.29e-03	4.36e-03	7.34e-02	2.19e-03	2.22e-03
Comb. score	1.61	7.39	-0.25	1.23	1.04
59	**Apoptosis_Homo sapiens_hsa04210**				
Overlap	11/140	13/140	10/140	18/140	15/140
Adj. p-value	2.36e-03	5.83e-04	1.32e-02	1.94e-07	1.71e-05
Comb. score	2.62	15.57	5.80	29.92	17.09
60	**Non-alcoholic fatty liver disease (NAFLD)_Homo sapiens_hsa04932**				
Overlap	11/151	11/151	7/151	16/151	13/151
Adj. p-value	4.18e-03	6.78e-03	1.76e-01	1.06e-05	3.81e-04
Comb. score	1.85	6.23	-0.83	19.29	5.17
61	**Hippo signaling pathway_Homo sapiens_hsa04390**				
Overlap	11/153	13/153	14/153	12/153	13/153
Adj. p-value	4.57e-03	1.10e-03	3.99e-04	1.49e-03	4.28e-04
Comb. score	0.95	11.26	14.77	2.61	3.71
62	**Focal adhesion_Homo sapiens_hsa04510**				
Overlap	13/202	6/202	9/202	12/202	9/202
Adj. p-value	4.86e-03	5.75e-01	1.49e-01	1.17e-02	1.15e-01
Comb. score	1.50	-1.12	-0.51	-0.21	-2.34
63	**FoxO signaling pathway_Homo sapiens_hsa04068**				
Overlap	10/133	8/133	9/133	11/133	14/133
Adj. p-value	5.12e-03	5.27e-02	2.25e-02	1.56e-03	3.59e-05
Comb. score	0.61	0.90	2.82	2.28	12.44
64	**Tight junction_Homo sapiens_hsa04530**				
Overlap	10/139	6/139	5/139	8/139	9/139
Adj. p-value	6.88e-03	2.54e-01	4.09e-01	4.54e-02	1.81e-02
Comb. score	0.18	-1.22	-1.56	-2.11	-1.68
65	**Phospholipase D signaling pathway_Homo sapiens_hsa04072**				
Overlap	10/144	4/144	7/144	7/144	8/144
Adj. p-value	8.76e-03	6.70e-01	1.50e-01	1.16e-01	5.33e-02
Comb. score	0.50	-1.06	-0.78	-2.41	-2.32

Appendix B

Clonal evolution in Multiple Myeloma

Supplementary Notes: Casewise Clonal evolution

Therapies given to patients-
V= Bortezomib
C= Cyclophosphamide
T= Thalidomide
R= Lenalidomide
D=Dexamethasone
M= Melphalan
P=Pomalidomide

Branching Clonal Evolution

SM0018 (Female / 54 years old / MM R-ISS2 with OS of 182.43, PFS of 106.14 weeks)

- Evolution pattern: Branching
- Total clones= 4; 2 founder(s) (3, 1); Rising clones (cellular prevalence at TP1 to TP2) = 3 (15.67 to 30.43), 1 (0.00 to 42.74)
- TMB at TP1 is 0.59; TMB at TP2 became 1.44
- Therapy: VCD
- Two founder clones (1,3) were detected at diagnosis that diversified into 4 clones by the time of progression (Figure B.1). The founder clone 3 possessed mutations in 5 genes including MUC6 (p.(Ala2054Val) that reduced in cellular prevalence from 0.31 at TP1 to 0.26 at TP2), rising mutations in NBPF1(c.3444G>A (p.(LysTer1148=)), NPIPB15 (p.(Ala238Thr)), GSTA2 (p.(Pro110Ser)), and MUC17 (p.(Thr959Ala)). Similarly, founder clone 1 had multiple mutations that increased with progression. These comprised of driver mutations in RNF213 (p.(Val1195Met)), KMT2C (p.(Tyr987His)) a tumor suppressor gene and others.

SM0022 (Male / 62 years old / MM R-ISS2 with OS of 304.86, PFS of 74.71 weeks)

- Evolution pattern: Branching
- Total clones= 4; 2 founder(s) (2, 4); Rising clones (cellular prevalence at TP1 to TP2) = 4 (0.00 to 45.15); Falling clones (cellular prevalence at TP1 to TP2) = 2 (31.76 to 23.14)
- TMB at TP1 is 1.75; TMB at TP2 became 1.43
- Therapy: VCD
- This patient had 4 clones of which clones 2 and 4 were founders at diagnosis (Figure B.2). Clone 2 had several falling mutations such as in CHI3L1 (5'UTR), ATP8B1(p.(Gln461Lys)), UNC80 (p.(Arg1030His)) and mutations with rising cellular prevalence such as termination in PABPC1 (p.(Glu345Ter)), LIMS1 (p.(Arg74His)) and others. Founder clone 4 had predominantly rising mutations in MYC (3'UTR), MTA1(splice variant), BAGE2 (5'UTR), AHNAK2 (p.(Met2187Val)), CARMIL3 (p.(Gly1161Val)) and others.

SM0024 (Male / 52 years old / MM R-ISS2 with OS of 121.29, PFS of 116.71 weeks)

- Evolution pattern: Branching
- Total clones= 3; 1 founder(s) (2); Rising clones (cellular prevalence at TP1 to TP2) = 2 (51.32 to 70.19)
- TMB at TP1 is 0.23; TMB at TP2 became 0.29
- Therapy: VCD, ASCT+VRD
- There were 3 clones in this patient that arose from a single founder clone 2. Numerous mutations emerged in this patient that evolved with rising cellular prevalence. Twelve mutations were detected and consisted of F5 (p.(His1327Arg)), KRT6B (p.(Ile365Val)), HERC1 (p.(Asp76Glu)), GOLGA6A(p.(Gln505Glu)), ZNF705G (p.(Gly47Arg)) and others.

SM0025 (Male / 48 years old / MM R-ISS2 with OS of 179.57, PFS of 127.43 weeks)

- Evolution pattern: Branching
- Total clones= 5; 2 founder(s) (2, 1); Rising clones (cellular prevalence at TP1 to TP2) = 2 (50.00 to 50.00); Falling clones (cellular prevalence at TP1 to TP2) = 1 (50.00 to 47.39)
- TMB at TP1 is 0.82; TMB at TP2 became 0.5
- Therapy: VCD Two founder clones (1,2) were enriched in CDKN1B (p.(Phe64Val)) and CSNK2A3 (p.(Arg280Gln)) in clone 1 that tended to decrease with progression and with stable levels of CACNA1G (p.(Leu1174Met)), MLF1 (p.(Ser102Tyr)) and RTP3 (p.(Asn74Ser)) in clone 2.

SM0067 (Female / 54 years old / MM R-ISS2 with OS of 70.14, PFS of 65.71 weeks)

- Evolution pattern: Branching
- Total clones= 5; 2 founder(s) (3, 2); Rising clones (cellular prevalence at TP1 to TP2) = 3 (28.66 to 36.80), 2 (0.00 to 42.50); ;
- TMB at TP1 is 0.2; TMB at TP2 became 1.42
- Therapy: VCD
- Founder clone 3 had abundant PDE4DIP (p.(His1598Arg)), CSTL1 (p.(Arg66Lys)) and FCRL6 (p.(Gln423Ter) rising mutations while reducing prevalence of KRTAP1-1 (p.(Pro58Arg)) and CSTL1 (p.(Arg66Lys)). The founder clone 2 had two rising mutations in OR4A16 (p.(Ile166Phe)) and MYH8 (p.(Lys638Asn)).

SM0082 (Male / 60 years old / MM R-ISS2 with OS of 275.86, PFS of 198.57 weeks)

- Evolution pattern: Branching
- Total clones= 4; 2 founder(s) (4, 1); Rising clones (cellular prevalence at TP1 to TP2) = 4 (22.66 to 44.09); Falling clones (cellular prevalence at TP1 to TP2) = 1 (32.52 to 0.00);
- TMB at TP1 is 0.82; TMB at TP2 became 0.77
- Therapy: VRD, ASCT
- This patient had two founder clones that branched to total 4 clones at the time of progression. Founder clone 1 had all falling mutations in USP31 (p.(Phe567Val)), METTL15 (p.(Asn31Lys)) and other genes, all of which diminished by progression. On the contrary, founder clone 4 had two mutations ZNF285 (p.(Lys292Glu)) and CCDC13 (p.(Arg25Trp)) that increased in prevalence further on progression.

SM0094 (Female / 44 years old / MM R-ISS2 with OS of 215.57, PFS of 67.86 weeks)

- Evolution pattern: Branching
- Total clones= 3; 1 founder(s) (3); Rising clones (cellular prevalence at TP1 to TP2) = 3 (41.72 to 59.76); ;
- TMB at TP1 is 2.42; TMB at TP2 became 0.58
- Therapy: RD, VCD_VD, CTD,VRD, CRP, PAD
- A single founder clone 3 was identified at diagnosis that evolved into three clones by TP2. The founder clone had multiple mutations in drivers such as FAT4 (rising p.(Ala807Val)) and PABPC1 (falling 3'UTR variant) and other rising mutations among LAMC1 (p.(Ile458Val)), NABP2 (p.(Glu118Lys)) and others.

SM0102 (Male / 48 years old / MM R-ISS3 with OS of 140.86, PFS of 45.00 weeks)

- Evolution pattern: Branching
- Total clones= 4; 3 founder(s) (3, 2, 4); Rising clones (cellular prevalence at TP1 to TP2) = 3 (33.33 to 33.33), 4 (0.00 to 21.39); Falling clones (cellular prevalence at TP1 to TP2) = 2 (20.47 to 0.00);
- TMB at TP1 is 88.1; TMB at TP2 became 91.85
- Therapy: VRD, VD
- Driver mutations in PDE4DIP (splice variant), tumor suppressor gene PIKR3 (p.(Asn329Lys)), oncogene CYP19A1, (3'UTR), MAP2K1 (splice variant), FANCA (p.(Gly501Ser)), SERPINB3 (p.(Gly351Ala)), PLCG1 (p.(Ile813Thr)), ALK (p.(Lys1491Arg)), DROSHA (3'UTR), PTCH1 (p.(Pro1315Leu)) were observed in clone 3. Clone 2 had NTRK1(p.(His604Tyr)), KRAS (3'UTR), PDPR (p.(Thr29Ala)), FAT1 (p.(Thr2261Met)), IL3 (p.(Pro27Ser)), PABPC1 (p.(Gly579Ser)) mutations. Founder clone 4 had NTRK1 (p.(Gly613Val)), PTPN14 (splice variant), EXO1 (p.(Glu589Lys)), BRCA1 (3'UTR) and other mutations.

SM0113 (Male / 61 years old / MM R-ISS2 with OS of 267.29, PFS of 267.29 weeks)

- Evolution pattern: Branching
- Total clones= 3; 2 founder(s) (2, 3); Rising clones (cellular prevalence at TP1 to TP2) = 2 (19.06 to 36.56); Falling clones (cellular prevalence at TP1 to TP2) = 3 (22.34 to 0.00);
- TMB at TP1 is 0.48; TMB at TP2 became 0.93
- Therapy: RD, VRD Representative mutations in founder clones 2 and 3 at diagnosis comprised of DNAH17 (p.(Arg879His)), LRP5 (splice variant), IRAK1 (p.(Asn345Ser)), ZNF98 (p.(Tyr350Cys)), ACOXL (p.(Thr255Met)), EXD3 (p.(Arg38Trp)) and others.

SM0152 (Male / 70 years old / MM R-ISS3 with OS of 238.00, PFS of 106.71 weeks)

- Evolution pattern: Branching
- Total clones= 4; 1 founder(s) (4); ; Falling clones (cellular prevalence at TP1 to TP2) = 4 (66.29 to 54.25);
- TMB at TP1 is 0.79; TMB at TP2 became 1.52
- Therapy: VTD-VD, CTD-RD A single founder clone 4 had mutations in driver HOXD13 (p.(Gly11Ala)), in MST1L (p.(Trp403Ter)), DHX58 (p.(Arg523Gln)), KLK14 (p.(Gln33Arg)) and others that diversified into total 4 clones by progression.

SM0172 (Female / 69 years old / MM R-ISS3 with OS of 51.86, PFS of 21.86 weeks)

- Evolution pattern: Branching

- Total clones= 6; 3 founder(s) (1, 4, 6); Rising clones (cellular prevalence at TP1 to TP2) = 1 (32.00 to 33.33), 6 (0.00 to 15.95); Falling clones (cellular prevalence at TP1 to TP2) = 4 (15.27 to 0.00);
- TMB at TP1 is 0.31; TMB at TP2 became 0.35
- Therapy: RD, VRD
- In this patient, three out of 6 clones were founders. Founder clone 1 had PLET1 (p.(Ser142Pro)), PABPC3 (p.(Met251Ile)), CCDC173 (5'UTR), clone 4 carried mutations in AK2 (3'UTR), SMARCB1 (p.(Val234Met)), CEL (p.(Ile488Thr)) whereas clone 6 had FCGBP (p.(Ala3916Val)), TTC30A (p.(Val446Ile)) and UGT1A5 (p.(Gly259Arg)).

SM0197 (Male / 45 years old / MM R-ISS3 with OS of 155.86, PFS of 50.00 weeks)

- Evolution pattern: Branching
- Total clones= 3; 1 founder(s) (1); Rising clones (cellular prevalence at TP1 to TP2) = 1 (45.78 to 50.31); ;
- TMB at TP1 was 2.19; TMB at TP2 became 1.39
- Therapy: VCD, VTD-DT,ASCT+VRD- RD
- Only one founder was detected at diagnosis. This clone 1 had 18 mutations, notably in drivers DIS3 (p.(Ile348Lys)), DICER1 (p.(Glu235Gly)), EPHA7 (p.(Met450Ile)), VWF (p.(Gly1922Ala)) and others.

SM0224 (Female / 61 years old / MM R-ISS2 with OS of 72.00, PFS of 72.00 weeks)

- Evolution pattern: Branching
- Total clones= 6; 3 founder(s) (5, 3, 4); Rising clones (cellular prevalence at TP1 to TP2) = 5 (33.33 to 33.33), 4 (0.00 to 33.33); Falling clones (cellular prevalence at TP1 to TP2) = 3 (33.33 to 0.00);
- TMB at TP1 is 48.4; TMB at TP2 became 23.23
- Therapy: VCD, RD
- Founder clone 5 had driver mutations in SUFU (c.1299T>C(p.(Ile433=))), PGR (p.(Ser344Thr)), CAMTA1 (p.(Asn1177Lys)), RHPN2 (p.(Gln384Arg)), SIRPA (p.(Gly75Ala)), BARD1 (p.(Arg378Ser)), ZNF292 (p.(Ile1740Val)). Clone 3 was characterized by mutations in PDE4DIP (p.(Arg171Lys)), ERCC5 (p.(Cys529Ser)), FANCM (p.(Ile1460Val)), PLCB4 (p.(Thr998Ala)), EP300 (p.(Ile997Val)), FAT4 (p.(Gly3526Asp)). Clone 4 had PDE4DIP (p.(Leu1272Phe)), CLIP (p.(Pro1403Leu)), EP400 (p.(Leu1741Gln)), TRAF3 (p.(Met129Thr)), BLM (p.(Pro868Leu), p.(Val1321Ile)) and other mutations.

SM0237 (Female / 52 years old / MM R-ISS2 with OS of 225.29, PFS of 225.29 weeks)

- Evolution pattern: Branching
- Total clones= 5; 2 founder(s) (3, 5); Rising clones (cellular prevalence at TP1 to TP2) = 3 (28.77 to 50.00); Falling clones (cellular prevalence at TP1 to TP2) = 5 (28.82 to 0.00);
- TMB at TP1 is 0.66; TMB at TP2 became 0.39
- Therapy: RD, VRD-VD, ASCT+VRD-RD
- Founder clone 5 had one major mutation PABPC3 (p.(Val119Phe) that reduced in prevalence on progression but the other founder clone 3 picked up prevalence of 4 mutations viz. RAB3GAP2 (p.(Asn570Ser)), TPTE (p.(Val68Asp)), RAB11FIP5 (p.(Arg461Trp)) and PBK (splice variant).

SM0267 (Male / 42 years old / MM R-ISS2 with OS of 235.14, PFS of 235.14 weeks)

- Evolution pattern: Branching
- Total clones= 2; 2 founder(s) (1, 2); Rising clones (cellular prevalence at TP1 to TP2) = 2 (0.00 to 15.74); Falling clones (cellular prevalence at TP1 to TP2) = 1 (14.87 to 0.00);
- TMB at TP1 is 0.96; TMB at TP2 became 0.24
- Therapy: RD
- On diagnosis, founder clone 2 had 3 mutations BAGE2 (p.(Arg106Gln)), EPHA5 (p.(Lys626Glu)) and PCDH12 (p.(Gln500His)). The other founder clone 1 had multiple mutations e.g., PADI4 (p.(Gly112Ala)), PRAMEF1 (p.(Trp98Arg)), AP3S2 (p.(Phe23Leu)), CGB7 (5'UTR) and others.

SM0294 (Male / 48 years old / MM R-ISS2 with OS of 241.71, PFS of 241.71 weeks)

- Evolution pattern: Branching
- Total clones= 4; 2 founder(s) (4, 2); Rising clones (cellular prevalence at TP1 to TP2) = 4 (29.98 to 36.03); Falling clones (cellular prevalence at TP1 to TP2) = 2 (50.00 to 0.00);
- TMB at TP1 is 0.31; TMB at TP2 became 0.63
- Therapy: VRD
- Founder clone 2 had mutations in KIAA0586 (p.(Leu1568Pro)) and GRK4 (p.(Arg65Leu)), both of which were lost by progression. Clone 4 became predominant by progression and most of mutations in clone 4 increased in cellular prevalence with time. These mutations were CDK11B (c.1959T>G(p.(Ala653=))), KRTAP4-11 (p.(Leu161Val)), TTC30A (p.(Val446Ile)), SENP2 (p.(Thr301Lys)), ZAN (p.(Ala2761Pro)) and NUTM2F (p.(Ala589Gly)).

SM0299 (Female / 53 years old / MM R-ISS2 with OS of 90.14, PFS of 40.86 weeks)

- Evolution pattern: Branching
- Total clones= 3; 1 founder(s) (1); ; Falling clones (cellular prevalence at TP1 to TP2) = 1 (77.14 to 36.45);
- TMB at TP1 is 0.26; TMB at TP2 became 0.17
- Therapy: RD
- This patient had a single founder clone 1 that had a falling mutation in SPRN (p.(Thr7Met)) and an almost constant maintained mutation in OBSCN (5'UTR).

SM0311 (Female / 52 years old / MM R-ISS2 with OS of 104.43, PFS of 94.00 weeks)

- Evolution pattern: Branching
- Total clones= 5; 1 founder(s) (3); ; Falling clones (cellular prevalence at TP1 to TP2) = 3 (100.00 to 90.80);
- TMB at TP1 is 0.79; TMB at TP2 became 0.14
- Therapy: VRD-VD, ASCT-VRD,VCD
- A single founder clone 3 had falling mutations in ALG10 (p.(Val19Ile)), GC-SAML (5'UTR) and a consistent KLH5 (splice mutation).

SM0329 (Female / 54 years old / MM R-ISS2 with OS of 69.14, PFS of 67.86 weeks)

- Evolution pattern: Branching
- Total clones= 4; 2 founder(s) (2, 4); Rising clones (cellular prevalence at TP1 to TP2) = 2 (29.58 to 31.59), 4 (0.00 to 34.69);
- TMB at TP1 is 1.39; TMB at TP2 became 4.53
- Therapy: VD, VTD
- Two founder clones 2 and 4 were identified at diagnosis. Clone 2 had driver mutations in TP53 (p.(Arg158His)), MCM3AP (p.(Trp502Ter)), EGR1 (p.(Ser62Asn)). Similarly, clone 4 was mutated in multiple drivers such as KRAS (p.(Gly13Asp)), DNMT1 (p.(Ala1334Thr)), CNOT3 (3'UTR), KMT2C (p.(Tyr987His)) etc.

SM0339 (Male / 71 years old / MM R-ISS2 with OS of 100.43, PFS of 28.86 weeks)

- Evolution pattern: Branching
- Total clones= 3; 2 founder(s) (2, 3); Rising clones (cellular prevalence at TP1 to TP2) = 3 (0.00 to 19.73); Falling clones (cellular prevalence at TP1 to TP2) = 2 (45.66 to 18.93);

- TMB at TP1 is 1.74; TMB at TP2 became 0.33
- Therapy: VRD, RD
- Founder clone 3 showed an increase in prevalence while the other founder clone 2 reduced in cellular prevalence with time. Clone 2 had mutations such as those in F5 (p.(Arg740Ter)), RBL2 (p.(Gln783Ter)), HIST1H4D (p.(Ala70Val)) etc. While clone 3 had driver mutations in PARP4 (splice variant), IGLL5 (p.(Pro20His)) and others.

SM0343 (Male / 60 years old / MM R-ISS2 with OS of 123.57, PFS of 121.57 weeks)

- Evolution pattern: Branching
- Total clones= 3; 2 founder(s) (1, 3); Rising clones (cellular prevalence at TP1 to TP2) = 1 (8.65 to 35.34); Falling clones (cellular prevalence at TP1 to TP2) = 3 (10.74 to 0.00);
- TMB at TP1 is 0.86; TMB at TP2 became 0.69
- Therapy: VD, RD+H3:H31
- Three rising mutations were found in founder clone 1 (STK36 (p.(Leu434Pro)), GPR160 (p.(Ile262Thr)), KLHL38 (p.(Ile334Val))) that increased in prevalence from diagnosis to progression. On the contrary, founder clone 3 had mutations that reduced relatively with time. These included FLG (p.(Trp3555Arg)), CFP (p.(Pro237His)), FBXW11 (p.(Arg356Ser)) and others.

SM0351 (Male / 58 years old / MM R-ISS2 with OS of 226.86, PFS of 108.86 weeks)

- Evolution pattern: Branching
- Total clones= 3; 1 founder(s) (1); ; Falling clones (cellular prevalence at TP1 to TP2) = 1 (71.36 to 68.79);
- TMB at TP1 is 1.13; TMB at TP2 became 1.44
- Therapy: RD
- One founder clone 1 was found with mutations that increased in prevalence (WDFY4 (p.(Leu841Met)), DIAPH3 (p.(Leu1034Ter)), BRAF (p.(Val600Glu))) or decreased with time (CABLES1 (p.(Lys496Arg)), CLOCK (p.(Ala400Gly)), KLHL31 (p.(Ala203Ser)) etc).

SM0370 (Female / 55 years old / MM R-ISS3 with OS of 228.00, PFS of 160.57 weeks)

- Evolution pattern: Branching
- Total clones= 4; 2 founder(s) (1, 2); Rising clones (cellular prevalence at TP1 to TP2) = 1 (21.95 to 25.48), 2 (0.00 to 26.65); ;

- TMB at TP1 is 0.27; TMB at TP2 became 0.57
- Therapy: RD, VRD-VD
- Driver mutations in BCL7A (p.(Arg4Gly)) in founder clone 1 and DOT1L(p.(Phe1474Tyr)), MLH1 (p.(Cys142Arg)) were observed in clone 2 among others.

SM0422 (Female / 45 years old / MM R-ISS2 with OS of 205.86, PFS of 146.00 weeks)

- Evolution pattern: Branching
- Total clones= 3; 2 founder(s) (1, 3); Rising clones (cellular prevalence at TP1 to TP2) = 3 (0.00 to 16.19); Falling clones (cellular prevalence at TP1 to TP2) = 1 (46.07 to 15.64);
- TMB at TP1 is 0.47; TMB at TP2 became 0.29
- Therapy: VRD, VD, VTD, VCD, CRP
- In this patient, founder clone 1 had falling mutations in FANK1 (p.(Gln4Ter)), DDX60L (p.(Cys336Tyr)) whereas clone 3 had rising mutations in CPED1 (p.(Ala551Gly), STAP2 (p.(Ala366Gly)), CETP (c.1212C>T(p.(Phe404=))) and others.

SM0433 (Male / 72 years old / MM R-ISS2 with OS of 186.29, PFS of 183.29 weeks)

- Evolution pattern: Branching
- Total clones= 3; 1 founder(s) (3); ; Falling clones (cellular prevalence at TP1 to TP2) = 3 (69.20 to 47.65);
- TMB at TP1 is 1.51; TMB at TP2 became 1.11
- Therapy: VD, VRD, RD
- An individual founder clone 3 had rising variations in driver NRAS (p.(Gln61Arg)), PRDM4 (p.(His99Arg)), RBM5 (p.(Gly759Arg), p.(Arg633Thr), p.(Ser744Ile)), MAPK10 (p.(Val244Leu)). Mutations that decreased in prevalence in this clone with time included HCAR1 (p.(Val277Met)), OBSCN (p.(Val634Met)), CD163L1 (p.(Gly1074Cys)) and others.

SM0505 (Male / 60 years old / MM R-ISS2 with OS of 168.71, PFS of 72.00 weeks)

- Evolution pattern: Branching
- Total clones= 5; 2 founder(s) (3, 5); Rising clones (cellular prevalence at TP1 to TP2) = 3 (18.24 to 42.96), 5 (0.00 to 50.00);
- TMB at TP1 is 0.77; TMB at TP2 became 1.97
- Therapy: VCD, VD

- Two founder clones were observed at diagnosis. Clone 3 had multiple mutations while clone 5 had 4 mutations that emerged before progression. The latter were BRINP2 (p.(Val134Gly)), FAN1 (p.(Arg581Ter)), COL14A1 (p.(Pro1717Arg)) and FSD1L (p.(Phe57Leu)). Clone 3 had TRPC6 (p.(Gly20Arg)), RORC (p.(Leu501Val)), HDAC10 (p.(Asn142Lys)) and others.

SM0510 (Male / 39 years old / MM R-ISS2 with OS of 48.86, PFS of 46.14 weeks)

- Evolution pattern: Branching
- Total clones= 3; 2 founder(s) (3, 1); Rising clones (cellular prevalence at TP1 to TP2) = 1 (0.00 to 32.58); Falling clones (cellular prevalence at TP1 to TP2) = 3 (16.67 to 0.00);
- TMB at TP1 is 0.15; TMB at TP2 became 0.78
- Therapy: VRD
- Founder clone 1 was a rising clone with mutations in driver CIC (p.(Glu125Ter)), FAM171A1 (p.(Thr303Met)), TRPV4 (p.(Gly20Arg)), DNAH1 (p.(Tyr1899Cys)), MUC16 (p.(Pro14112His)), MLLT6 (p.(Pro45Thr)) etc. Cellular prevalence of mutations e.g., in PARP4 (5'UTR), SYNM (p..(Ala212Val)) and EXOC7 (3'UTR) were observed to fall with time in clone 3.

SM0546 (Female / 54 years old / MM R-ISS2 with OS of 143.71, PFS of 100.29 weeks)

- Evolution pattern: Branching
- Total clones= 2; 2 founder(s) (1, 2); Rising clones (cellular prevalence at TP1 to TP2) = 2 (0.00 to 50.00); Falling clones (cellular prevalence at TP1 to TP2) = 1 (17.54 to 0.00);
- TMB at TP1 is 0.55; TMB at TP2 became 0.03
- Therapy: VRD, RD
- Founder Clone 2 had a single mutation in SF1 (p.(Pro64Ser)). Whereas founder clone 1 had several mutations in drivers CR1 (p.(Glu888Asp)), PTPRS (p.(Arg1798Cys)), BCORL1 (p.(Ile175Asn)), EGR1 (p.(Asn61Lys)), HIST1H1D (p.(Asn78Ser), p.(Ser87Arg)), FAM3C (p.(Ser75Gly)) and others.

SM0581 (Female / 58 years old / MM R-ISS2 with OS of 138.14, PFS of 18.86 weeks)

- Evolution pattern: Branching
- Total clones= 4; 1 founder(s) (4); ; Falling clones (cellular prevalence at TP1 to TP2) = 4 (93.11 to 90.22);

- TMB at TP1 is 0.55; TMB at TP2 became 0.63
- Therapy: PD
- This patient had a single founder clone with a mutation in MYRFL (p.(Ser157Ala)) and in PPIAL4G (3'UTR).

SM0584 (Male / 67 years old / MM R-ISS2 with OS of 64.00, PFS of 51.00 weeks)

- Evolution pattern: Branching
- Total clones= 5; 2 founder(s) (4, 2); Rising clones (cellular prevalence at TP1 to TP2) = 2 (0.01 to 47.04); Falling clones (cellular prevalence at TP1 to TP2) = 4 (32.58 to 21.86);
- TMB at TP1 is 2.47; TMB at TP2 became 1.54
- Therapy: RD
- Founder clone 4 had multiple mutations such as MTA2 (p.(Pro184Ala)), MUC5AC (p.(Gly1085Ser)), UNG (p.(Ala264Thr)), NCAPD2 (p.(Thr1331Ala)). The other founder clone 2 had mutations in NM1 (p.(Ser16Leu)) and others.

SM0588 (Male / 53 years old / MM R-ISS2 with OS of 97.57, PFS of 23.00 weeks)

- Evolution pattern: Branching
- Total clones= 3; 1 founder(s) (2); ; Falling clones (cellular prevalence at TP1 to TP2) = 2 (77.14 to 45.18);
- TMB at TP1 is 65.79; TMB at TP2 became 46.64
- Therapy: VRD
- Multiple driver mutations were present in a single founder clone 2. These consisted of actionable WRN (p.(Leu1074Phe)), ROS1 (splice variant), MAP3K1 (p.(Asp806Asn)), FBN2 (p.(Pro2784Leu), p.(Met2311Val)), ATXN7 (p.(Lys264Arg), p.(Val862Met)), ATR (p.(Arg2425Gln)), CUL3 p.(Val573Ile), FOXD4L1 (p.(Asn162Lys)) and others.

SM0660 (Male / 63 years old / MM R-ISS2 with OS of 166.71, PFS of 49.14 weeks)

- Evolution pattern: Branching
- Total clones= 4; 3 founder(s) (3, 1, 4); Rising clones (cellular prevalence at TP1 to TP2) = 4 (0.00 to 8.89); Falling clones (cellular prevalence at TP1 to TP2) = 3 (10.05 to 7.29), 1 (33.33 to 0.00);
- TMB at TP1 is 0.26; TMB at TP2 became 0.18
- Therapy: VRD, VD

- Of the 3 founder clones at diagnosis, only one clone 4 increased in cellular prevalence by progression and had mutations in LMAN2L (p.(Arg255Cys)), TIPARP (p.(Glu370Lys)) and WDFY3 (p.(Arg941Met)). The other clone 3 had LEPR (p.(Pro266Ser)) and NDST3 (p.(Lys498Thr)) mutations while the clone 1 had 3' UTR mutations in ASCC1, GIT2, CDH24 etc.

SM0678 (Male / 68 years old / MM R-ISS3 with OS of 133.00, PFS of 112.43 weeks)

- Evolution pattern: Branching
- Total clones= 4; 3 founder(s) (3, 4, 2); Rising clones (cellular prevalence at TP1 to TP2) = 3 (10.16 to 14.82), 2 (0.00 to 17.66); Falling clones (cellular prevalence at TP1 to TP2) = 4 (18.82 to 0.00);
- TMB at TP1 is 0.38; TMB at TP2 became 1.34
- Therapy: VCD, VD
- This patient had three founder clones at the time of diagnosis. Clone 3 had driver mutations in MAX (p.(Arg33Ter)), clone 4 in SLC45A3 (5'UTR), GREM1 (3'UTR), and clone 2 in PLCB4 (p.(Ile222Val)), FANCD2 (p.(Tyr632Cys)) and FGFR3(p.(Arg671Gly)).

SM0686 (Male / 63 years old / MM NA with OS of 141.00, PFS of 84.71 weeks)

- Evolution pattern: Branching
- Total clones= 3; 1 founder(s) (2); ; Falling clones (cellular prevalence at TP1 to TP2) = 2 (53.70 to 40.38);
- TMB at TP1 is 0.76; TMB at TP2 became 0.41
- Therapy: VCD, VD
- Only single founder clone was identified in this patient and had a mutation in tumor suppressor gene KLF2 (p.(Leu104Pro)), and in other genes PBX1 (p.(Tyr384Ter)) and MKNK2 (p.(Lys4Asn)).

SM0698 (Male / 44 years old / MM R-ISS2 with OS of 149.14, PFS of 54.29 weeks)

- Evolution pattern: Branching
- Total clones= 3; 1 founder(s) (2); Rising clones (cellular prevalence at TP1 to TP2) = 2 (51.63 to 68.78); ;
- TMB at TP1 is 0.48; TMB at TP2 became 1.12
- Therapy: VCD, VD
- One founder clone was present at TP1 with multiple mutations in RIF1 (p.(Leu645His)), C6orf118 (p.(Lys327Met)), and many more.

SM0726 (Male / 56 years old / MM R-ISS3 with OS of 126.14, PFS of 64.00 weeks)

- Evolution pattern: Branching
- Total clones= 3; 2 founder(s) (1, 3); Rising clones (cellular prevalence at TP1 to TP2) = 3 (0.00 to 18.19); Falling clones (cellular prevalence at TP1 to TP2) = 1 (50.00 to 0.00);
- TMB at TP1 is 0.19; TMB at TP2 became 0.35
- Therapy: VRD, VD
- Founder clone 5 had 5'UTR variations in ACTA2, STK39 and EPHB1 genes while clone 3 had mutations in PIK3C2A (p.(Asn1003Ser)), C1orf167 (p.(Gly1188Ser)), SPINK5 (p.(Arg711Gln)) and others.

SM0738 (Male / 61 years old / MM NA with OS of 143.43, PFS of 80.00 weeks)

- Evolution pattern: Branching
- Total clones= 5; 1 founder(s) (2); Rising clones (cellular prevalence at TP1 to TP2) = 2 (100.00 to 100.00); ;
- TMB at TP1 is 0.85; TMB at TP2 became 1.1
- Therapy: VRD, RD
- A single founder dominated at diagnosis with mutations in APOA4 (p.(Arg220Cys)), TENM4 (p.(Arg2298Trp)), PCDHGC3 (p.(Val701Gly)) and 5'UTR variant in OSTF1 gene.

SM0740 (Male / 47 years old / MM R-ISS2 with OS of 124.57, PFS of 51.71 weeks)

- Evolution pattern: Branching
- Total clones= 3; 1 founder(s) (1); Rising clones (cellular prevalence at TP1 to TP2) = 1 (49.10 to 94.36); ;
- TMB at TP1 is 2.01; TMB at TP2 became 2.6
- Therapy: VRD, VD
- This patient also had a single founder clone carrying driver mutations in ENPEP (3'UTR), and PIM1 (p.(His6Leu), p.(Thr287Pro)) genes.

SM0755 (Male / 46 years old / MM R-ISS2 with OS of 126.86, PFS of 100.29 weeks)

- Evolution pattern: Branching
- Total clones= 3; 1 founder(s) (1); Rising clones (cellular prevalence at TP1 to TP2) = 1 (45.20 to 54.62); ;
- TMB at TP1 is 1.74; TMB at TP2 became 1.91

- Therapy: VCD, VRD
- One founder clone was detected at TP1 with several driver mutations such as MAX (p.(Arg35Cys)), FAT1 (p.(Phe3823Val)), PIM1 (p.(Ser74Ala)), RECQL4 (p.(Ser886Arg)) and NR4A3 (p.(Glu591Lys)).

SM0779 (Female / 70 years old / MM R-ISS2 with OS of 129.57, PFS of 85.14 weeks)

- Evolution pattern: Branching
- Total clones= 2; 2 founder(s) (2, 1); Rising clones (cellular prevalence at TP1 to TP2) = 1 (0.00 to 32.99); Falling clones (cellular prevalence at TP1 to TP2) = 2 (13.43 to 0.00);
- TMB at TP1 is 0.48; TMB at TP2 became 1.74
- Therapy: VRD, RD
- At TP1, the founder clone 2 had mutations in EIF!AD (p.(Ala164Thr)), TAS2R43 (p.(Gly160Arg)), DVL3 (p.(Ser233Leu)) etc while the founder clone 1 had somatic mutations in driver genes such as KRAS (p.(Gln61His)), STAG2 (p.(Arg252Trp)) and CTNNB1 (p.(Asp162Glu)).

SM0815 (Male / 62 years old / MM R-ISS2 with OS of 78.29, PFS of 36.29 weeks)

- Evolution pattern: Branching
- Total clones= 3; 1 founder(s) (3); ; Falling clones (cellular prevalence at TP1 to TP2) = 3 (44.95 to 33.44);
- TMB at TP1 is 2.97; TMB at TP2 became 1.02
- Therapy: VRD, VD
- Only one founder clone carrying mutations in MYH2 (p.(Glu1940Lys)), GAL3ST1 (p.(Arg354His)), PLXND1 (3'UTR), GRM4 (p.(Arg351Cys)) and TREML1 (p.(Pro269Leu)) was present at diagnosis.

SM1288 (Female / 48 years old / MM R-ISS2 with OS of 46.57, PFS of 26.71 weeks)

- Evolution pattern: Branching
- Total clones= 3; 1 founder(s) (2); ; Falling clones (cellular prevalence at TP1 to TP2) = 2 (68.21 to 46.28);
- TMB at TP1 is 1.09; TMB at TP2 became 0.86
- Therapy: VCD
- This patient possessed single founder clone with mutations in drivers TP53 (p.(Cys277Phe)), DCC (p.(Leu334Ter)) and BRAF (p.(Val600Glu)).

SM1547 (Male / 34 years old / MM R-ISS2 with OS of 60.29, PFS of 57.14 weeks)

- Evolution pattern: Branching
- Total clones= 3; 1 founder(s) (2); Rising clones (cellular prevalence at TP1 to TP2) = 2 (54.63 to 55.53); ;
- TMB at TP1 is 1.88; TMB at TP2 became 1.67
- Therapy: VCD
- A single founder clone was identified with multiple mutations in NXPE1 (p.(Thr117Pro)), MMP26 (splice variant), NOS1 (p.(Arg904Gly)) and others.

SM007 (Female/58 years old/MM R-ISS3 with OS of 221 weeks, PFS of 175 weeks)

- Evolution pattern: Branching
- Total clones= 9; 3 founder(s) (3, 8, 9); Rising clones (cellular prevalence at TP1 to TP2) = 3 (33.33 to 33.33); Falling clones (cellular prevalence at TP1 to TP2) = 8 (33.33 to 27.41), 9 (33.33 to 0.00);
- TMB at TP1 is 134.43; TMB at TP2 became 96.61
- Therapy: VCD; PCD
- Among several somatic mutations present in founder clone 1, an actionable driver mutation in NRAS (p.(Gly12Ala)) and another in FOXD4L1 (p.(Arg145Cys)) were noticed. Additional somatic mutations observed in founder clone 2 included KIAA0556 (p.(Ser368Asn)), MIPEP (p.(Ser368Asn)) and others.

SM0052 (Male / 59 years old/MM R-ISS2 with OS of 242.29 weeks, PFS of 148.00 weeks)

- Evolution pattern: Branching
- Total clones= 5; 3 founder(s) (3, 4, 5); Rising clones (cellular prevalence at TP1 to TP2) = 3 (0.00 to 32.75); Falling clones (cellular prevalence at TP1 to TP2) = 5 (32.99 to 32.64), 4 (32.11 to 0.00);
- TMB at TP1 is 132.12; TMB at TP2 became 119.99
- Therapy: DT, ASCT
- This patient has several driver mutations such as those in FAM186A (p.(Leu1233Pro), p.(Lys187Gln)), ZNF626 (p.(Lys180Asn)), MERTK (p.(Ile518Val)), IL3 (p.(Pro27Ser)) in founder clone 5, in PTPRC (p.(Asp543Asn)), KRAS (3'UTR), BRCA2 (p.(Asn289His)), CYP19A1 (p.(Arg264Cys)) in founder 4, and in NCOR2 (p.(Ala1699Thr)), KMT2B (p.(Asp2364Gly)) in founder clone 3.

Linear evolution

SM0076 (Male / 72 years old / MM R-ISS2 with OS of 280.71, PFS of 252.71 weeks)

- Evolution pattern: Linear
- Total clones= 2; 2 founder(s) (2, 1); Rising clones (cellular prevalence at TP1 to TP2) = 2 (17.13 to 20.99), 1 (0.00 to 23.15); ;
- TMB at TP1 is 0.35; TMB at TP2 became 2.16
- Therapy: MPT
- Founder clone 2 carried mutations in PNPLA2 (p.(Ser170Ala)), MIXL1 (p.(Ala81Thr)) etc while founder clone 1 had mutations in actionable driver oncogene NRAS (p.(Gly12Ala)), FOXD4L1 (p.(Arg145Cys)) and many more.

SM0133 (Male / 50 years old / MM R-ISS2 with OS of 260.00, PFS of 260.00 weeks)

- Evolution pattern: Linear
- Total clones= 2; 2 founder(s) (2, 1); Rising clones (cellular prevalence at TP1 to TP2) = 2 (0.31 to 29.50), 1 (0.00 to 49.93); ;
- TMB at TP1 is 1.36; TMB at TP2 became 105.47
- Therapy: VTD-VD
- Founder clone 2 had driver mutations such as TSG SUFU (p.(Arg280Gln)), actionable oncogene RET (p.(Lys989Arg)), TSG TET1 (p.(Ser193Thr), p.(Ala256Val)), NOTCH2 (p.(Asn2008Ser), p.(Asp1327Gly)), KRAS (3'UTR), FLT1 (p.(Lys337Gln)), TP53BP1 (p.(Lys1141Gln)), BRCA1 (p.(Ser1613Gly), p.(Lys1183Arg), p.(Glu1038Gly)), DNMT1 (p.(Ile327Val)) while founder clone 1 had numerous mutations in drivers such as CLIP1 p.(Asp1080Glu), MLH3 (p.(Pro844Leu)), STIL (p.(Ala86Val)) and CIITA (p.(Leu45Val)).

SM0138 (Male / 69 years old / MM R-ISS2 with OS of 181.14, PFS of 85.14 weeks)

- Evolution pattern: Linear
- Total clones= 3; 1 founder(s) (1); Rising clones (cellular prevalence at TP1 to TP2) = 1 (82.81 to 97.65); ;
- TMB at TP1 is 79.67; TMB at TP2 became 80.72
- Therapy: RD
- One single founder clone was observed at diagnosis, which contained mutations in PDE4DIP (p.(Trp560Ter), p.(Arg295His)), PRRX1 (p.(Ser200Arg)), IKBKE (p.(Pro713Leu)), PTPN11 (p.(Asn18Ser)) and others.

SM0143 (Female / 43 years old / MM R-ISS2 with OS of 253.14, PFS of 139.71 weeks)

- Evolution pattern: Linear
- Total clones= 3; 2 founder(s) (3, 2); Rising clones (cellular prevalence at TP1 to TP2) = 3 (49.23 to 49.91), 2 (0.00 to 49.69); ;
- TMB at TP1 is 95; TMB at TP2 became 39.41
- Founder clone 3 had FANCA (p.(Gly809Asp)), MSH3 (p.(Ala1045Thr)), WRN(p.(Leu1074Phe)) founder clone 2 had TET1 (p.(Asp162Gly)), RBM15 (p.(Asn798Ser)), FAM46C (p.(His67Gln)) and additional driver mutations.

SM0208 (Male / 54 years old / MM R-ISS2 with OS of 179.00, PFS of 132.71 weeks)

- Evolution pattern: Linear
- Total clones= 4; 2 founder(s) (3, 4); Rising clones (cellular prevalence at TP1 to TP2) = 4 (0.00 to 50.00); Falling clones (cellular prevalence at TP1 to TP2) = 3 (22.68 to 17.85);
- TMB at TP1 is 0.18; TMB at TP2 became 0.55
- Therapy: VRD-VD, ASCT+VRD-RD, CPT, DCEP
- Founder clone 3 was a falling clone with GUCY1A2(p.(Cys725Tyr)), TTF2 (p.(Lys167Glu)), whereas clone 4 was a rising founder clone with ZNF778 (p.(Tyr701Cys)), SLC35G4 (p.(Leu45Met)), and other mutations.

SM0559 (Female / 31 years old / MM R-ISS2 with OS of 163.57, PFS of 22.00 weeks)

- Evolution pattern: Linear
- Total clones= 2; 2 founder(s) (2, 1); Rising clones (cellular prevalence at TP1 to TP2) = 1 (0.00 to 19.21); Falling clones (cellular prevalence at TP1 to TP2) = 2 (17.43 to 0.05);
- TMB at TP1 is 0.16; TMB at TP2 became 0.37
- Therapy: VCD
- The founder clone 2 had mutations with falling prevalence (SDK2 (p.(Ala1499Gly)), ST8SIA3 (p.(Ala45Thr))) whereas founder clone 1 had mutations rising prevalence before progression (MUC5B (p.(Pro2830Leu)), MYH4 (p.(Ile1106Met)), etc.)

SM0664 (Male / 47 years old / MM R-ISS2 with OS of 101.71, PFS of 28.29 weeks)

- Evolution pattern: Linear
- Total clones= 2; 2 founder(s) (1, 2); Rising clones (cellular prevalence at TP1 to TP2) = 2 (0.00 to 18.56); Falling clones (cellular prevalence at TP1 to TP2) = 1 (30.07 to 0.97);

- TMB at TP1 is 4.69; TMB at TP2 became 1.16
- Therapy: VRD, CTD, DCEP
- In this patient, founder clone 1 had driver mutations in FAT3 (p.(Ser3322Arg)), SPRTN (p.(Val183Asp)), MGA (p.(Ser1263Ter)), EP300 (p.(Asn607Thr)) while founder clone 2 had mutations in KMT2C (p.(Lys339Asn)), KMT5A (p.(Pro60Leu)), TRIM60 (p.(Trp44Arg)) and others.

SM0667 (Female / 65 years old / MM R-ISS2 with OS of 130.57, PFS of 89.29 weeks)

- Evolution pattern: Linear
- Total clones= 4; 1 founder(s) (4); Rising clones (cellular prevalence at TP1 to TP2) = 4 (82.50 to 91.33); ;
- TMB at TP1 is 0.24; TMB at TP2 became 0.96
- Therapy: VRD, CRD
- A single founder clone was present in this patient at TP1 and harboured mutations in LRRC378 (p.(Gly652Arg)) and POM121 (p.(Pro478Leu)).

SM1595 (Female / 58 years old / MM R-ISS3 with OS of 95.86, PFS of 94.71 weeks)

- Evolution pattern: Linear
- Total clones= 2; 2 founder(s) (2, 1); Rising clones (cellular prevalence at TP1 to TP2) = 1 (0.00 to 21.40); Falling clones (cellular prevalence at TP1 to TP2) = 2 (20.47 to 10.83);
- TMB at TP1 is 0.91; TMB at TP2 became 3.14
- Therapy: VRD-VD, ASCT+VRD-RD
- This patient had mutations in founder clone 2 (such as PRAMEF1 (p.(Leu105Ter), (p.(Glu110Gly), CUL9 (p.(Glu377Ter))) while in founder clone 1 in KRAS (p.(Gly12Asp)), LRIG3 (p.(Val251Ile)), NBEA (p.(Arg2083His)) and others.

Stable with loss of clone

SM0115 (Female / 67 years old / MM R-ISS2 with OS of 220.00, PFS of 168.14 weeks)

- Evolution pattern: Stable with loss of clone
- Total clones= 2; 2 founder(s) (2, 1); ; Falling clones (cellular prevalence at TP1 to TP2) = 2 (33.84 to 0.58), 1 (50.00 to 0.00);

- TMB at TP1 is 54.83; TMB at TP2 became 0.53
- Therapy: RD, DT, CTD
- Founder clone 2 carried driver mutations in PDE4DIP (p.(Glu2001Gly)), PTPN14 (p.(Ile924Val)), HERC2 (p.(Val3327Met)), IGLL5 (p.(Thr211Ala)),, KMT2C (p.(Ala1685Ser), p.(Cys391Ter)). The other founder clone 1 had driver mutations in PGR (p.(Gln553Glu)), ATM (p.(His1380Tyr)), PRDM2 (p.(Ile586Thr)) , NUMA1 (p.(Ala794Gly)), KRAS (3'UTR), EXO1 (p.(Glu589Lys)), DIS3 (p.(Asn269Ser)) and others.

SM0167 (Female / 66 years old / MM R-ISS3 with OS of 67.57, PFS of 34.00 weeks)

- Evolution pattern: Stable with loss of clone
- Total clones= 2; 1 founder(s) (1); ; Falling clones (cellular prevalence at TP1 to TP2) = 1 (100.00 to 0.17);
- TMB at TP1 is 78.34; TMB at TP2 became 1.09
- Therapy: VRD, RD
- This patient had a single founder clone harbouring mutations in driver TSG TET1 (p.(Asp162Gly)), PDE4DIP (p.(Arg1978His), NUMA1 (p.(Glu809Asp), FAT3 (p.(Asn2293Ser)), CR1 (p.(Pro1827Arg)), TSG ATXN2 (p.(Ser248Asn)) and others.

SM0185 (Male / 63 years old / MM R-ISS2 with OS of 259.14, PFS of 147.00 weeks)

- Evolution pattern: Stable with loss of clone
- Total clones= 2; 2 founder(s) (2, 1); ; Falling clones (cellular prevalence at TP1 to TP2) = 2 (31.33 to 0.72), 1 (50.00 to 0.00);
- TMB at TP1 is 75.88; TMB at TP2 became 0.67
- Therapy: MP, MPT, RD

SM0266 (Male / 63 years old / MM R-ISS1 with OS of 184.14, PFS of 184.14 weeks)

- Evolution pattern: Stable with loss of clone
- Total clones= 2; 2 founder(s) (2, 1); Rising clones (cellular prevalence at TP1 to TP2) = 2 (0.92 to 29.01); Falling clones (cellular prevalence at TP1 to TP2) = 1 (16.69 to 0.00);
- TMB at TP1 is 0.63; TMB at TP2 became 0.47
- Therapy: VRD, VD Driver mutations in founder clone 2 included TSG TET1 (p.(Asn1018Ser)), CLIP1 (p.(Asp1080Glu)), PITRB (p.(Asp633Glu)), DNMT3B (c.1674T>C(p.(Tyr558=))), TSG FANCD2 (p.(Asn405Ser)), KMT2C (p.(Ala1685Ser)). Founder clone 1 had CPEB3 (p.(Ala499Gly)), PDE4DIP (p.(Val1371Ile)), LRP5 (p.(Val667Met)), FAT3(p.(Gln1726Arg)) and others.

SM0353 (Female / 61 years old / MM R-ISS3 with OS of 202.43, PFS of 133.43 weeks)

- Evolution pattern: Stable with loss of clone
- Total clones= 2; 2 founder(s) (1, 2); Rising clones (cellular prevalence at TP1 to TP2) = 1 (14.09 to 17.57); Falling clones (cellular prevalence at TP1 to TP2) = 2 (18.61 to 0.00);
- TMB at TP1 is 1.08; TMB at TP2 became 0.34
- Therapy: VRD-VD, VCD-VD, CPT
- Founder clone 1 possessed driver mutations at TP1 (MYO5A (p.(Arg90Ser))), similarly, founder clone 2 had CACNA1D (p.(Ser1224Tyr)), FAT1 (p.(Val43Met)), NSD2 (p.(Met397Ile)) and BRAF (p.(Val600Glu)) actionable driver mutations.

SM0471 (Female / 48 years old / MM R-ISS3 with OS of 220.43, PFS of 52.14 weeks)

- Evolution pattern: Stable with loss of clone
- Total clones= 3; 2 founder(s) (3, 1); Rising clones (cellular prevalence at TP1 to TP2) = 3 (25.46 to 27.33); Falling clones (cellular prevalence at TP1 to TP2) = 1 (33.07 to 0.00);
- TMB at TP1 is 1.13; TMB at TP2 became 0.32
- Therapy: RD, CTD, DT
- In this case, at diagnosis, founder clone 3 was loaded with actionable driver mutations in BIRC3 (p.(Tyr31His)), etc. The other founder clone 1 also had actionable driver mutations in BARD1 (p.(Tyr87His)), FGFR3 (p.(Cys275Tyr)) and others including IGLL5 (p.(Gln22Ter)), TNFAIP3 (p.(Arg45Ter)) and HIST1H1E (p.(Ala116Val)).

SM0808 (Male / 70 years old / MM R-ISS3 with OS of 83.14, PFS of 31.00 weeks)

- Evolution pattern: Stable with loss of clone
- Total clones= 2; 2 founder(s) (2, 1); Rising clones (cellular prevalence at TP1 to TP2) = 2 (7.46 to 22.78); Falling clones (cellular prevalence at TP1 to TP2) = 1 (15.81 to 0.00);
- TMB at TP1 is 2.44; TMB at TP2 became 0.69
- Therapy: VCD, VTD
- On diagnosis, two driver somatic mutations were observed in founder clone 2 (NBEA (p.(Ser200Cys)), IRF4 (p.(Lys123Arg))) and a few in founder clone 1 (ARID5B (p.(Asn640Ser)), DIS3 (p.(Arg780Lys)), AXIN2 (p.(Arg841Gln)), actionable BRAF (p.(Val600Glu)) and others).

SM0145 (Male / 60 years old/MM R-ISS3 with OS of 251.14 weeks, PFS of 123.29 weeks)

- Evolution pattern: Stable with loss of clone
- Total clones= 2; 2 founder(s) (1, 2); Falling clones (cellular prevalence at TP1 to TP2) = 2 (34.89 to 1.28), 1 (50.00 to 0.00);
- TMB at TP1 is 126.3; TMB at TP2 became 0.97
- Therapy: VRD-VD, VCD-RD
- Driver mutations found in founder clone 2 included PARP4 (p.(His490Gln)), MAX (p.(Met1?)), actionable MLH3 (p.(Pro844Leu)), HERC2 (p.(Val3327Met)), ALK (p.(Glu588Ala)), NOTCH4 (p.(Gly534Ser), p.(Lys117Gln)) and others. The founder clone also possessed driver mutations such as TCF7L2 (p.(Pro495Ala), RET (p.(Gly691Ser), CCDC6 (p.(Pro470Thr), ARID5B (p.(Asn299Lys)), ATM (p.(Leu263Pro), p.(Ser707Pro)), KMT2A (p.(Arg3564Trp)), NOTCH2 (p.(His1160Arg)).

Supplementary Figures: Casewise Clonal evolution

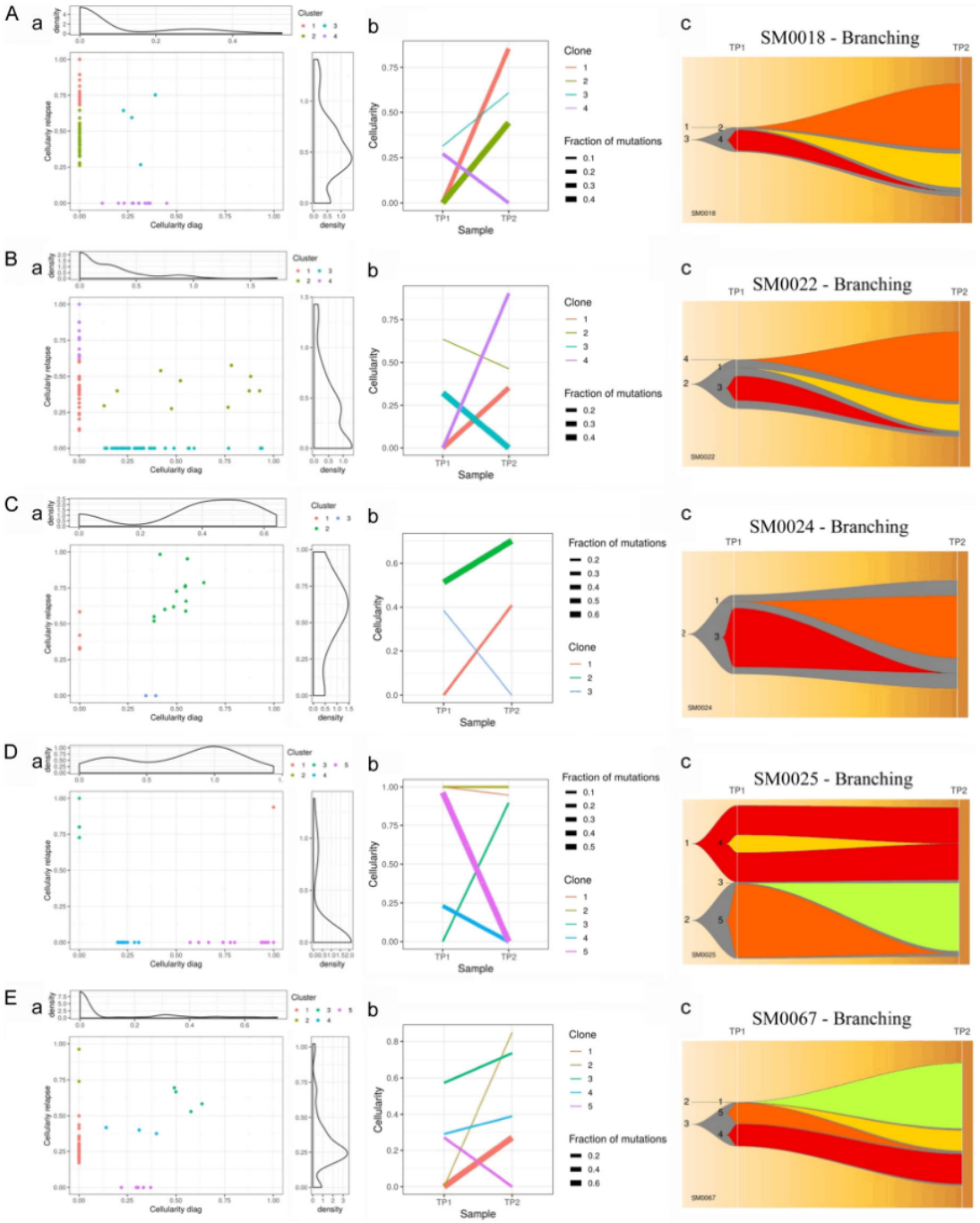


Figure B.1: (A-E) Clonal evolution in each case of MM. Representation of clonal evolution through (a) Density, (b) Evolution and (c) Fish plots across individual MM patients with branching, linear and stable with loss of clone patterns of clonal evolution.

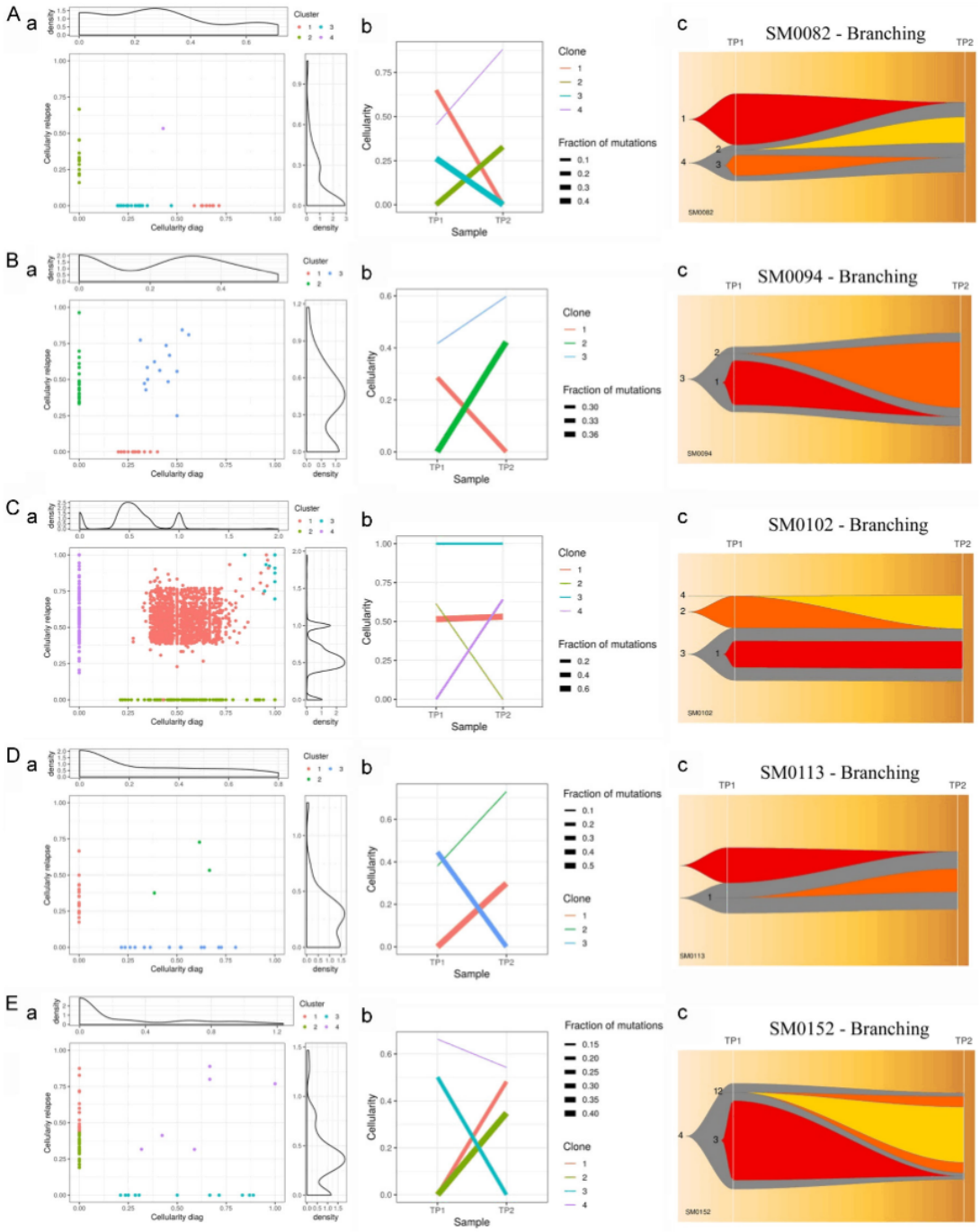


Figure B.2: (A-E) Clonal evolution in each case of MM. Representation of clonal evolution through (a) Density, (b) Evolution and (c) Fish plots across individual MM patients with branching, linear and stable with loss of clone patterns of clonal evolution.

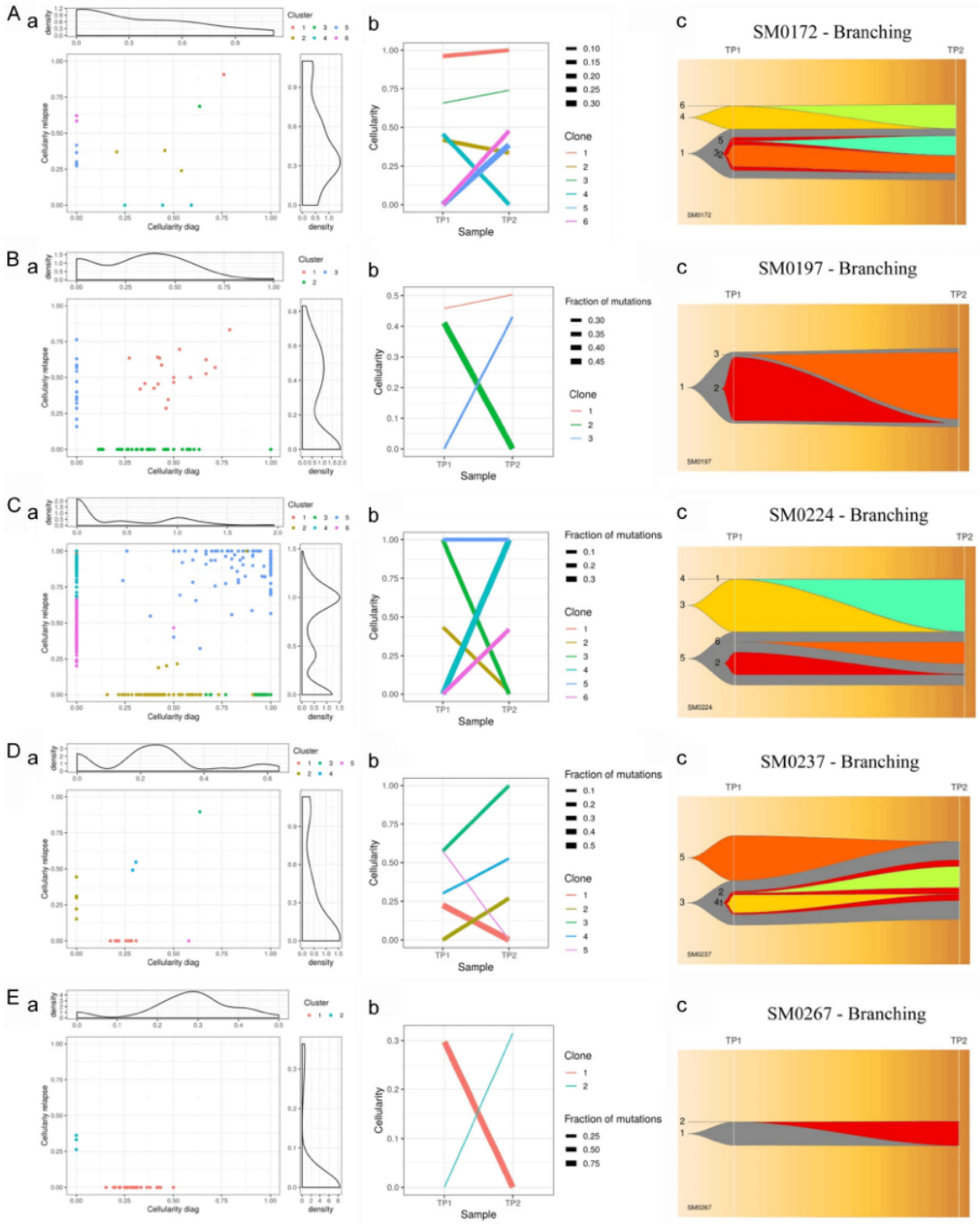


Figure B.3: (A-E) Clonal evolution in each case of MM. Representation of clonal evolution through (a) Density, (b) Evolution and (c) Fish plots across individual MM patients with branching, linear and stable with loss of clone patterns of clonal evolution.

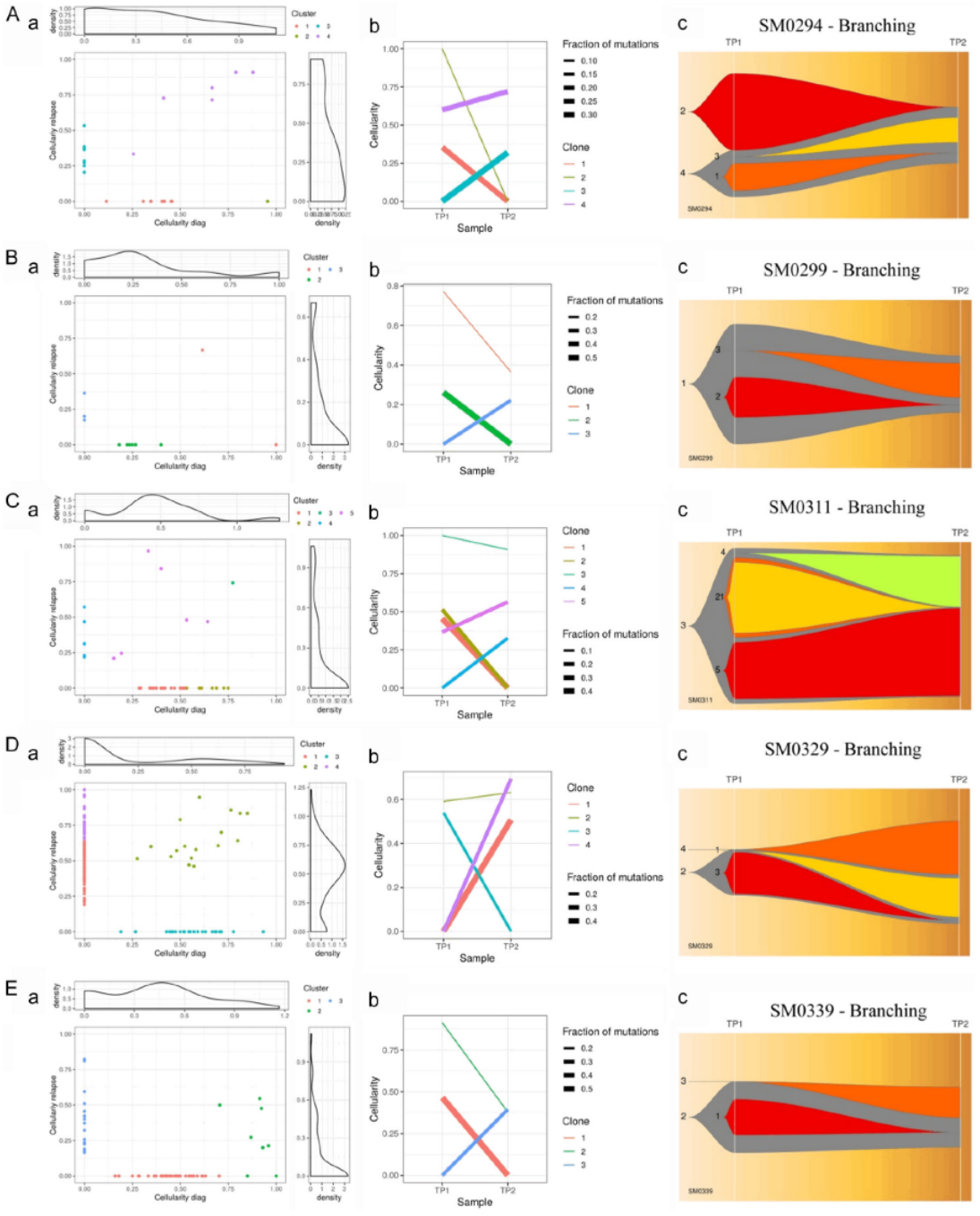


Figure B.4: (A-E) Clonal evolution in each case of MM. Representation of clonal evolution through (a) Density, (b) Evolution and (c) Fish plots across individual MM patients with branching, linear and stable with loss of clone patterns of clonal evolution.

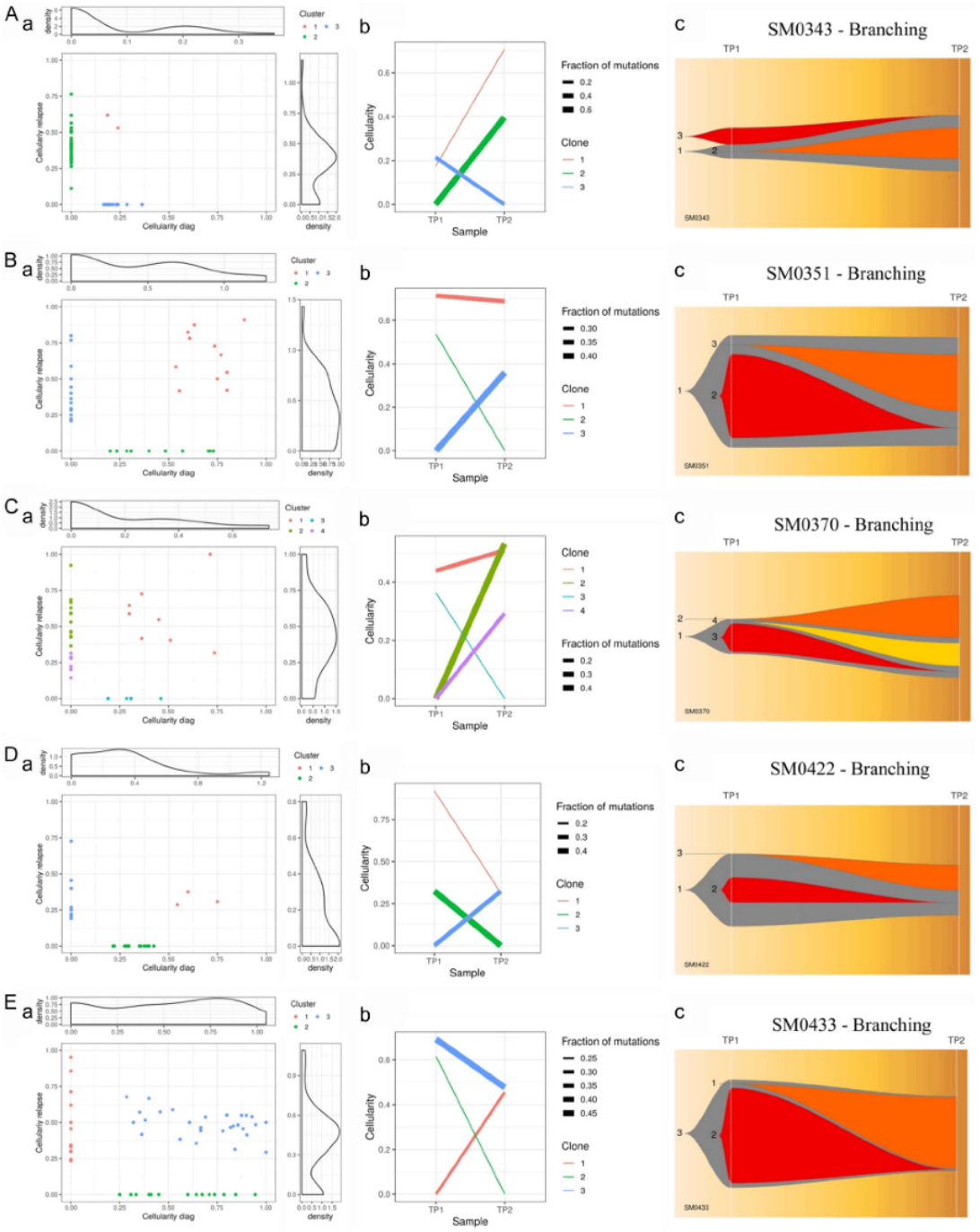


Figure B.5: (A-E) Clonal evolution in each case of MM. Representation of clonal evolution through (a) Density, (b) Evolution and (c) Fish plots across individual MM patients with branching, linear and stable with loss of clone patterns of clonal evolution.

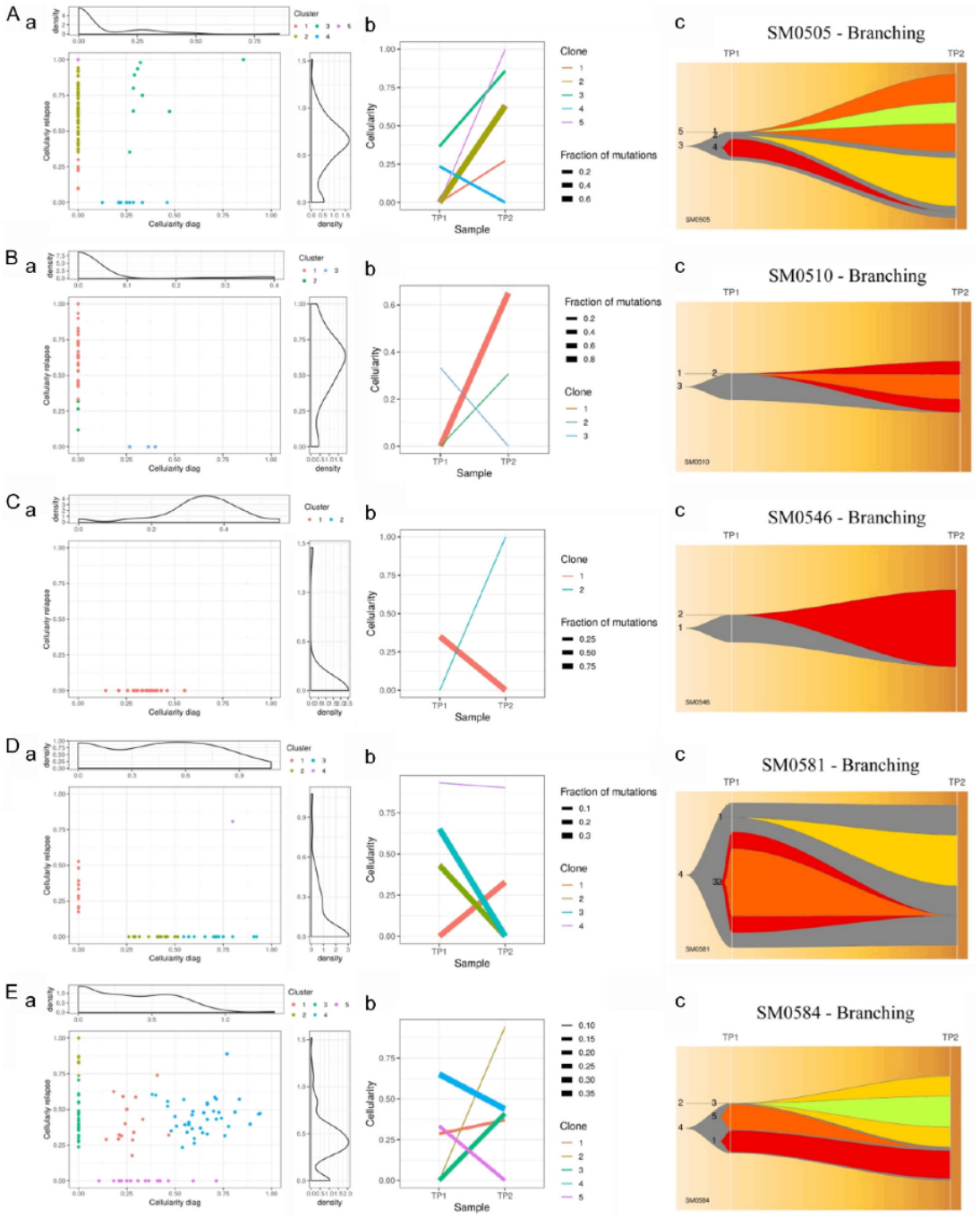


Figure B.6: (A-E) Clonal evolution in each case of MM. Representation of clonal evolution through (a) Density, (b) Evolution and (c) Fish plots across individual MM patients with branching, linear and stable with loss of clone patterns of clonal evolution.

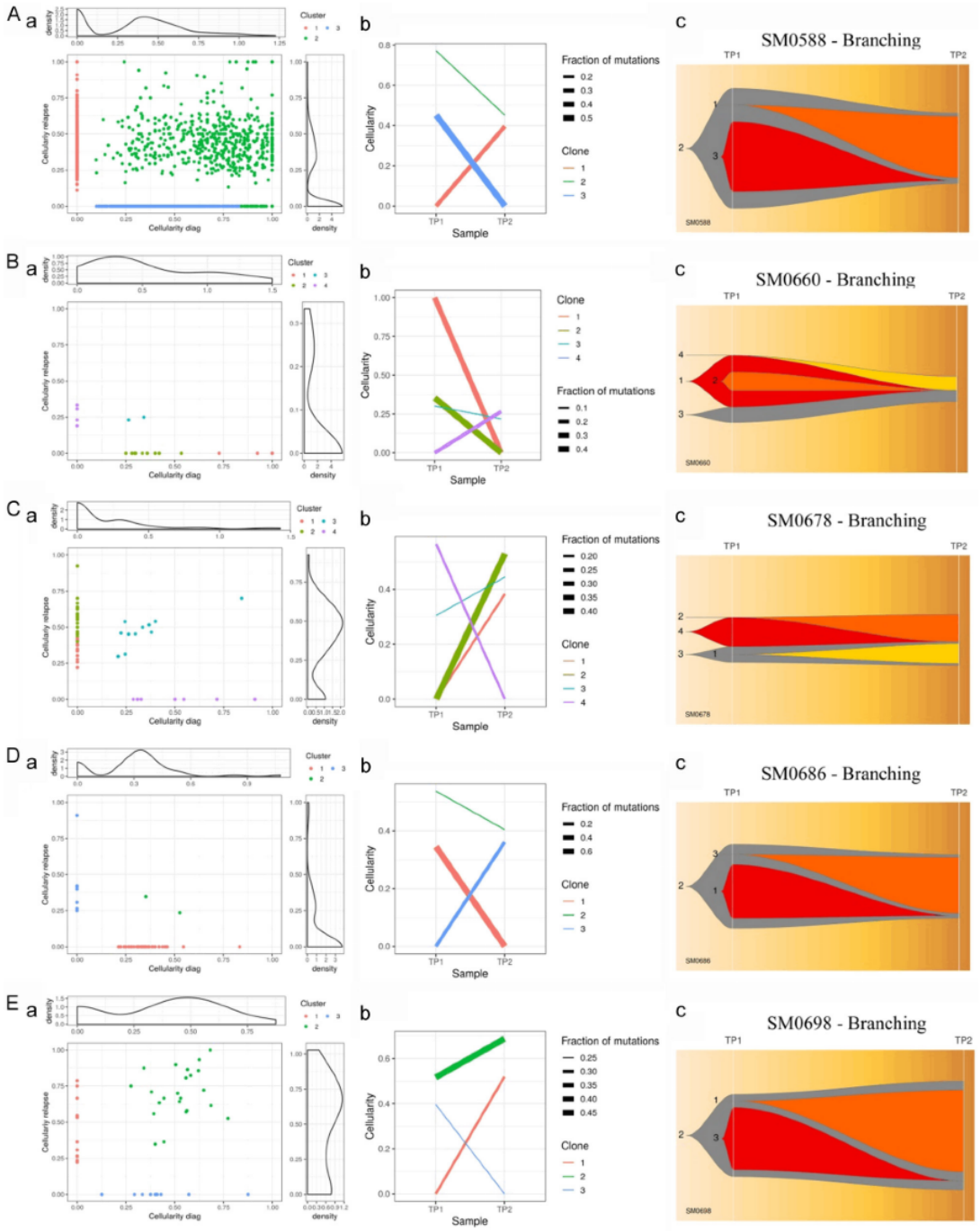


Figure B.7: (A-E) Clonal evolution in each case of MM. Representation of clonal evolution through (a) Density, (b) Evolution and (c) Fish plots across individual MM patients with branching, linear and stable with loss of clone patterns of clonal evolution.

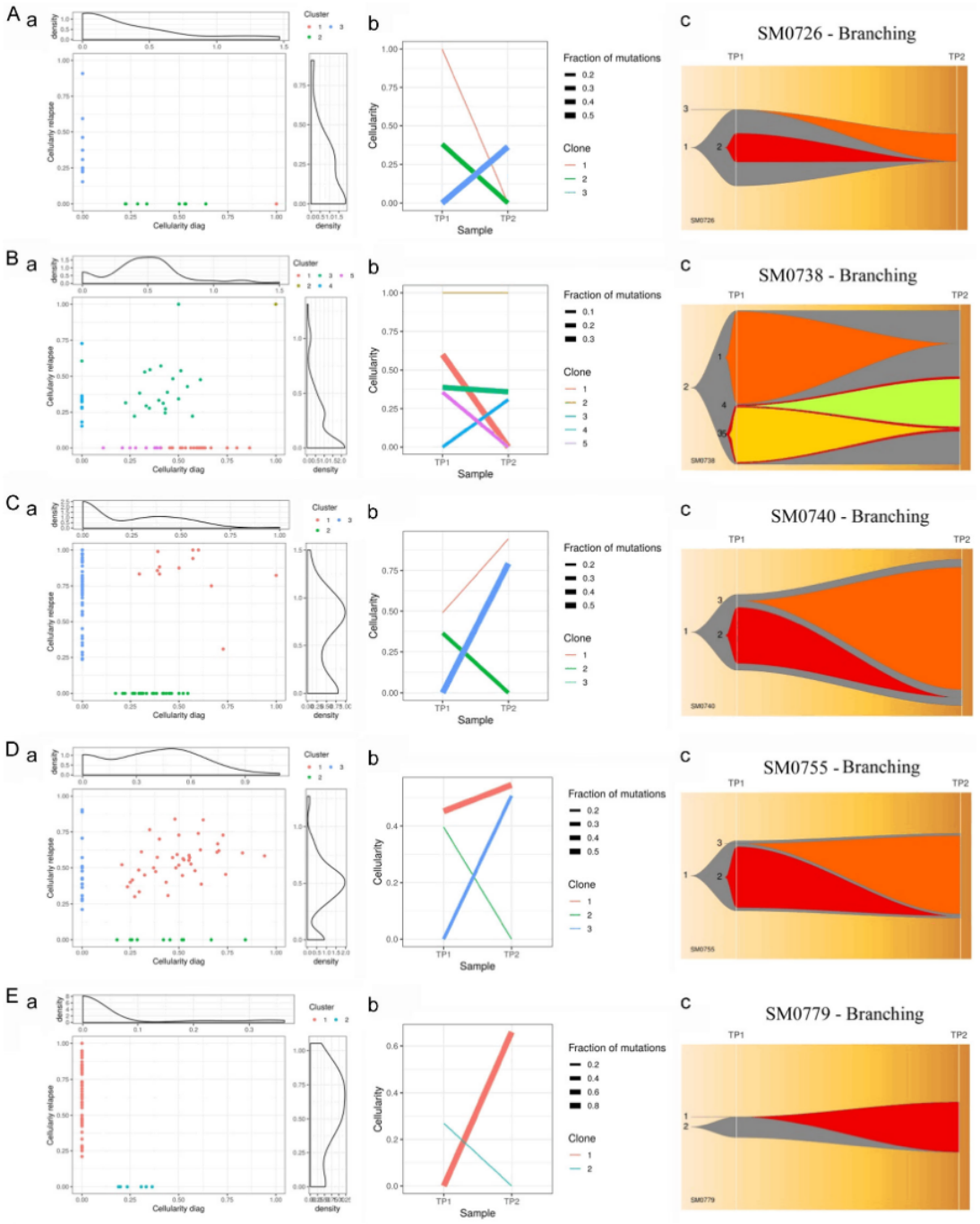


Figure B.8: (A-E) Clonal evolution in each case of MM. Representation of clonal evolution through (a) Density, (b) Evolution and (c) Fish plots across individual MM patients with branching, linear and stable with loss of clone patterns of clonal evolution.

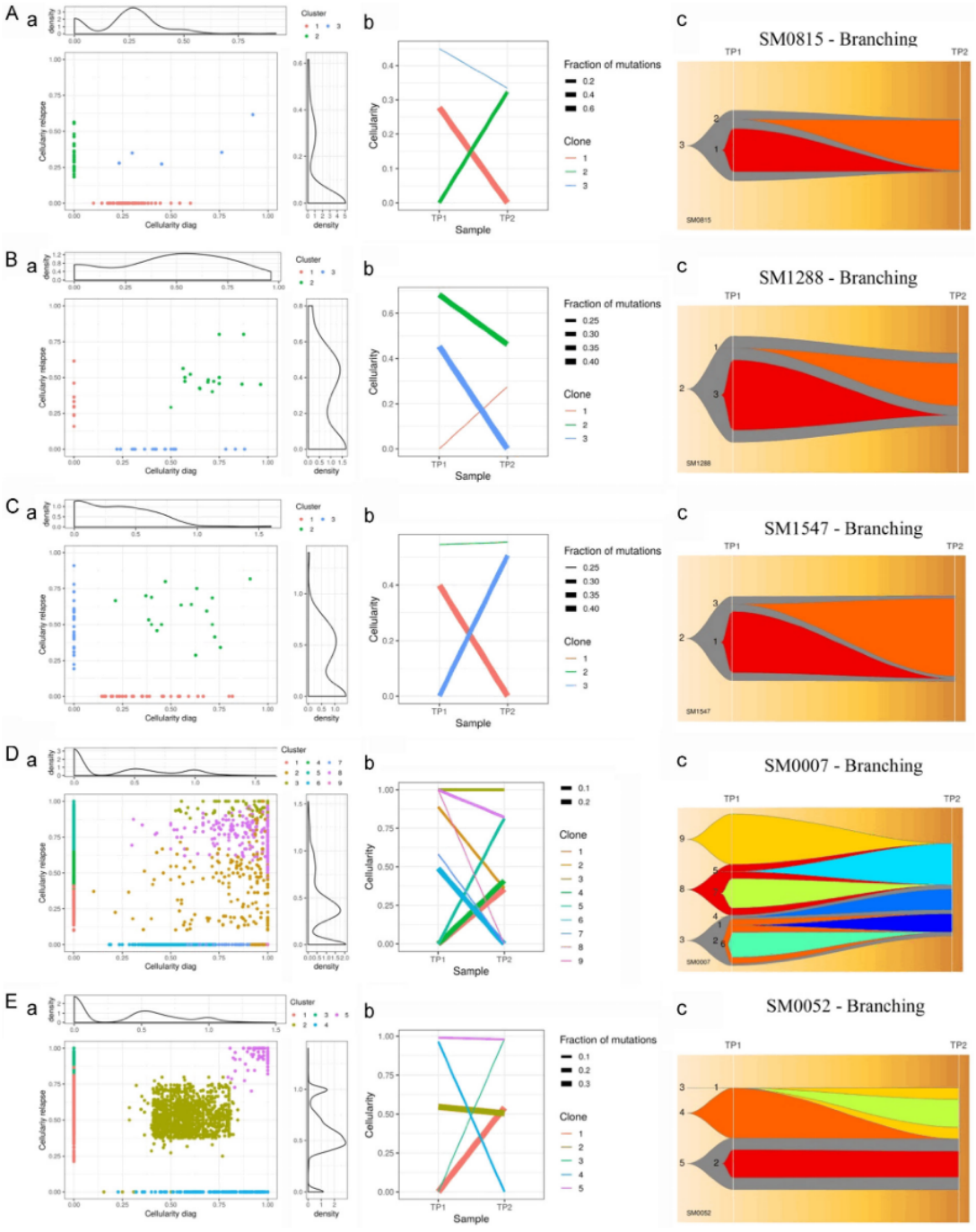


Figure B.9: (A-E) Clonal evolution in each case of MM. Representation of clonal evolution through (a) Density, (b) Evolution and (c) Fish plots across individual MM patients with branching, linear and stable with loss of clone patterns of clonal evolution.

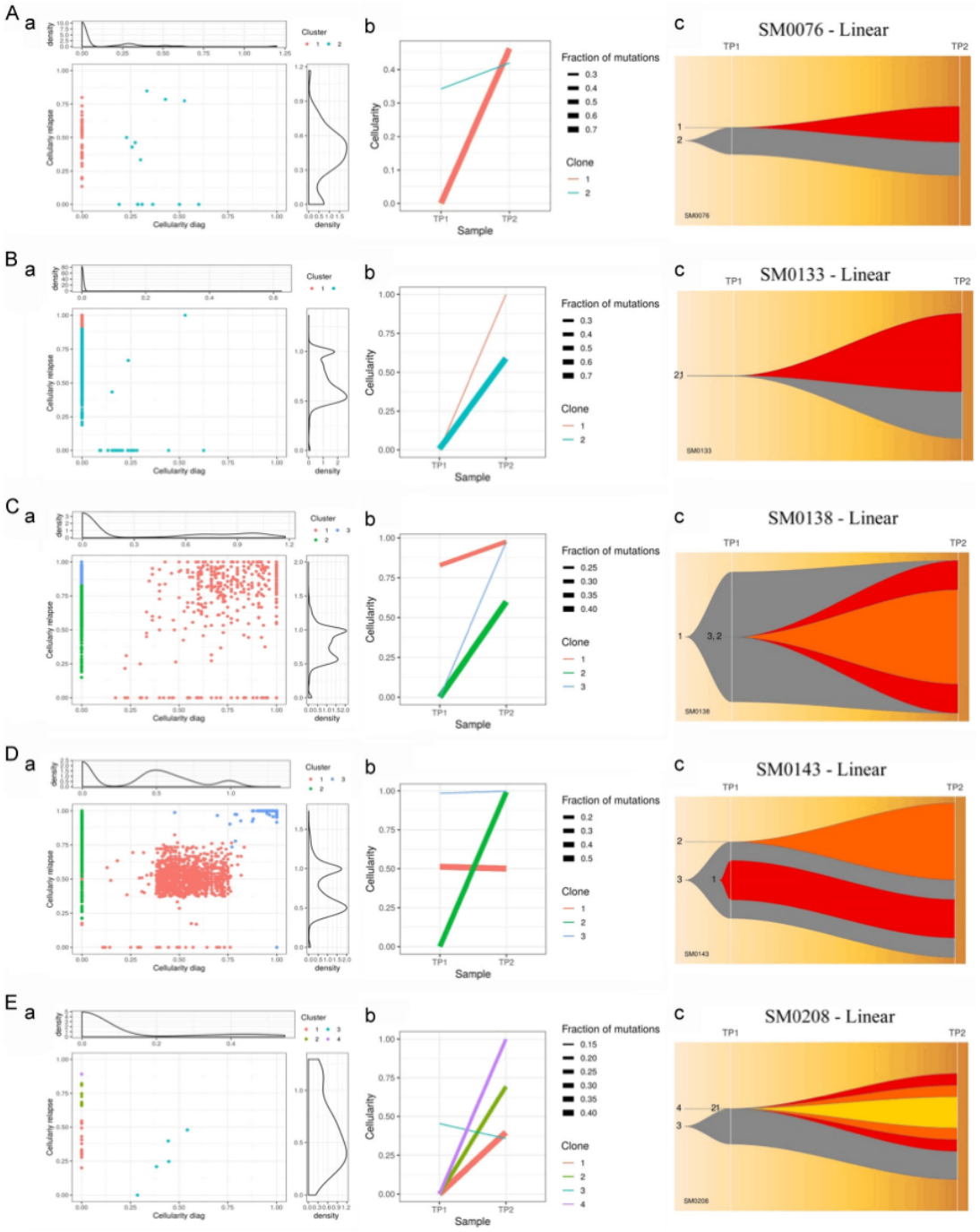


Figure B.10: (A-E) Clonal evolution in each case of MM. Representation of clonal evolution through (a) Density, (b) Evolution and (c) Fish plots across individual MM patients with branching, linear and stable with loss of clone patterns of clonal evolution.

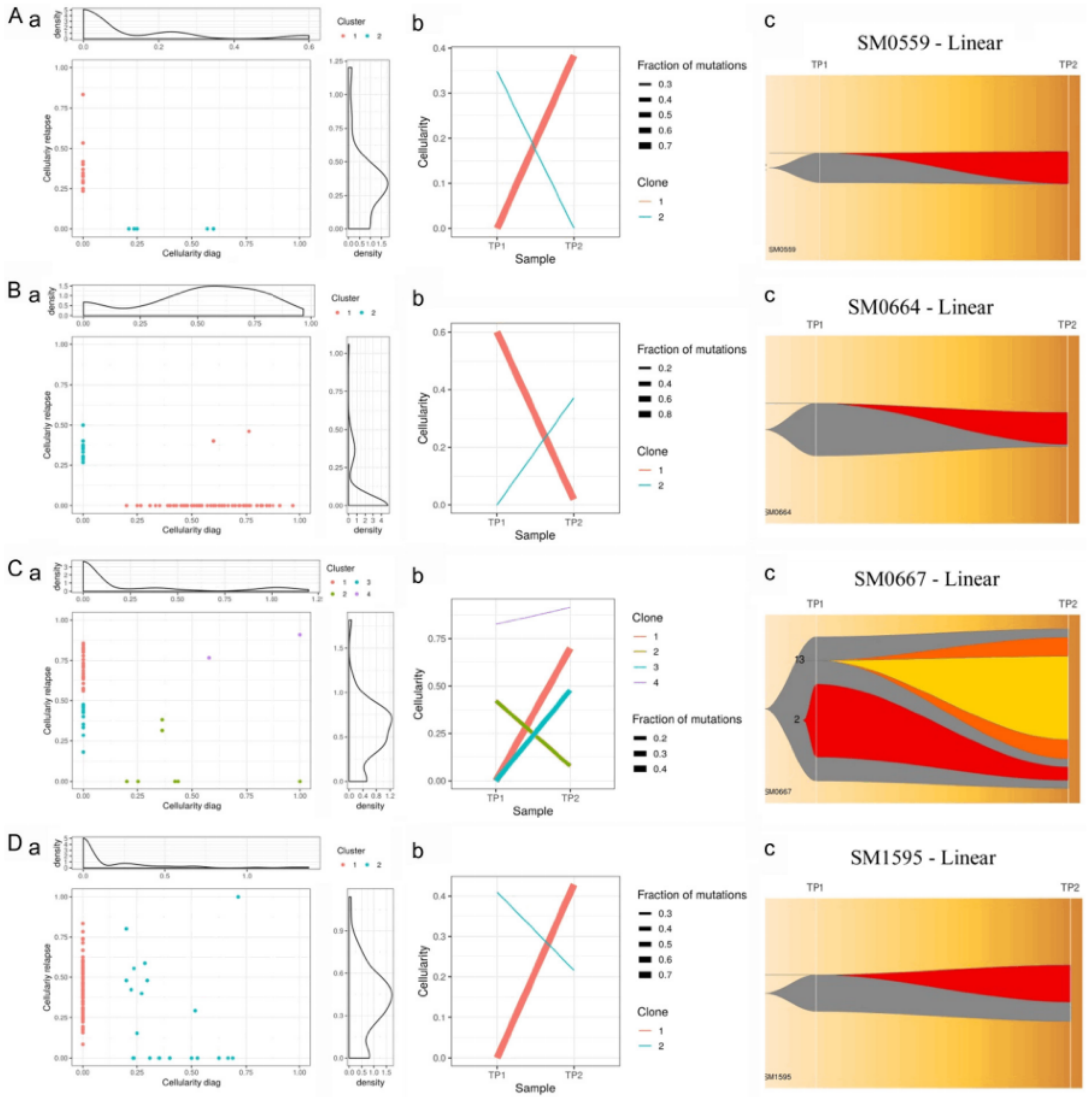


Figure B.11: (A-E) Clonal evolution in each case of MM. Representation of clonal evolution through (a) Density, (b) Evolution and (c) Fish plots across individual MM patients with branching, linear and stable with loss of clone patterns of clonal evolution.

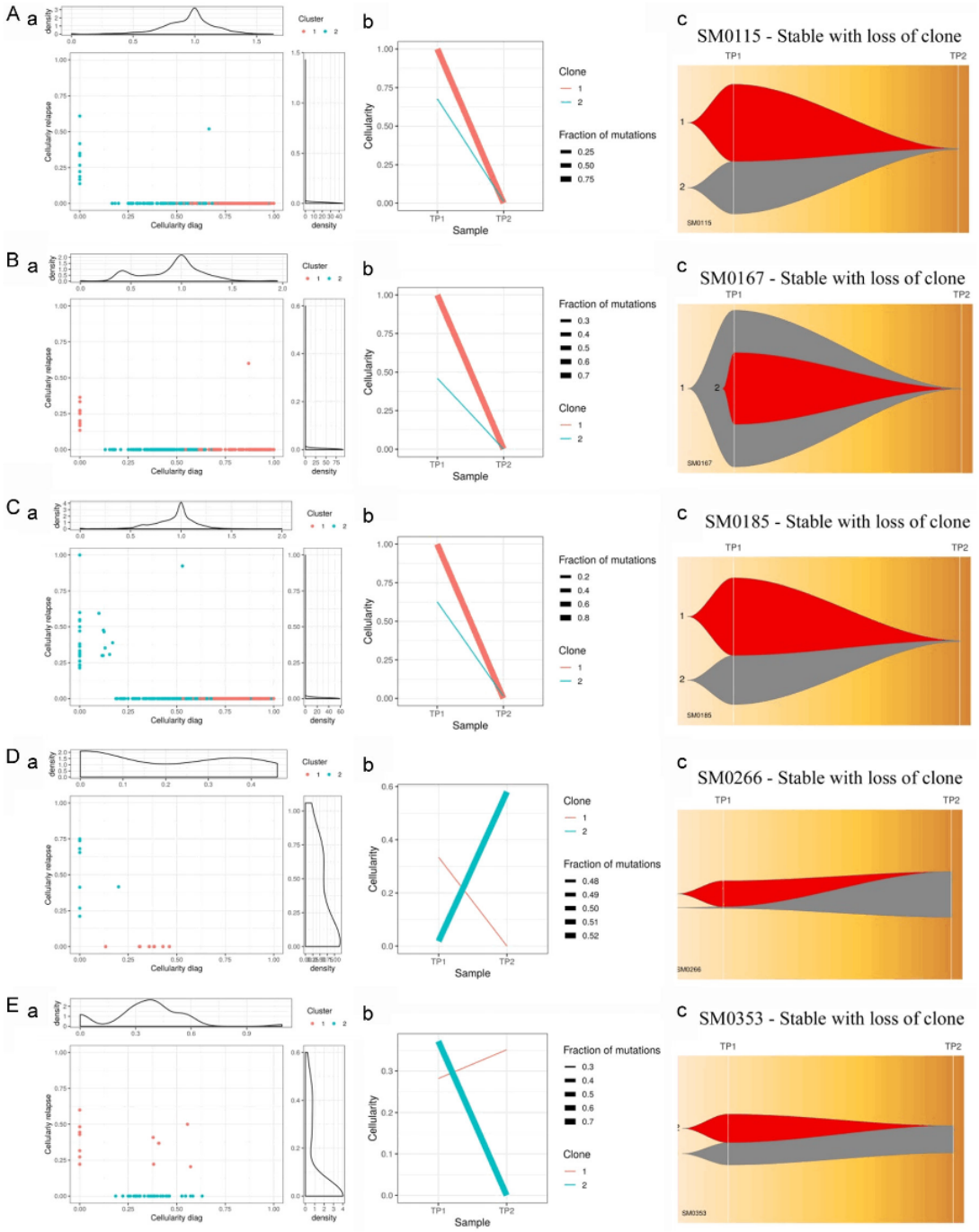


Figure B.12: (A-E) Clonal evolution in each case of MM. Representation of clonal evolution through (a) Density, (b) Evolution and (c) Fish plots across individual MM patients with branching, linear and stable with loss of clone patterns of clonal evolution.

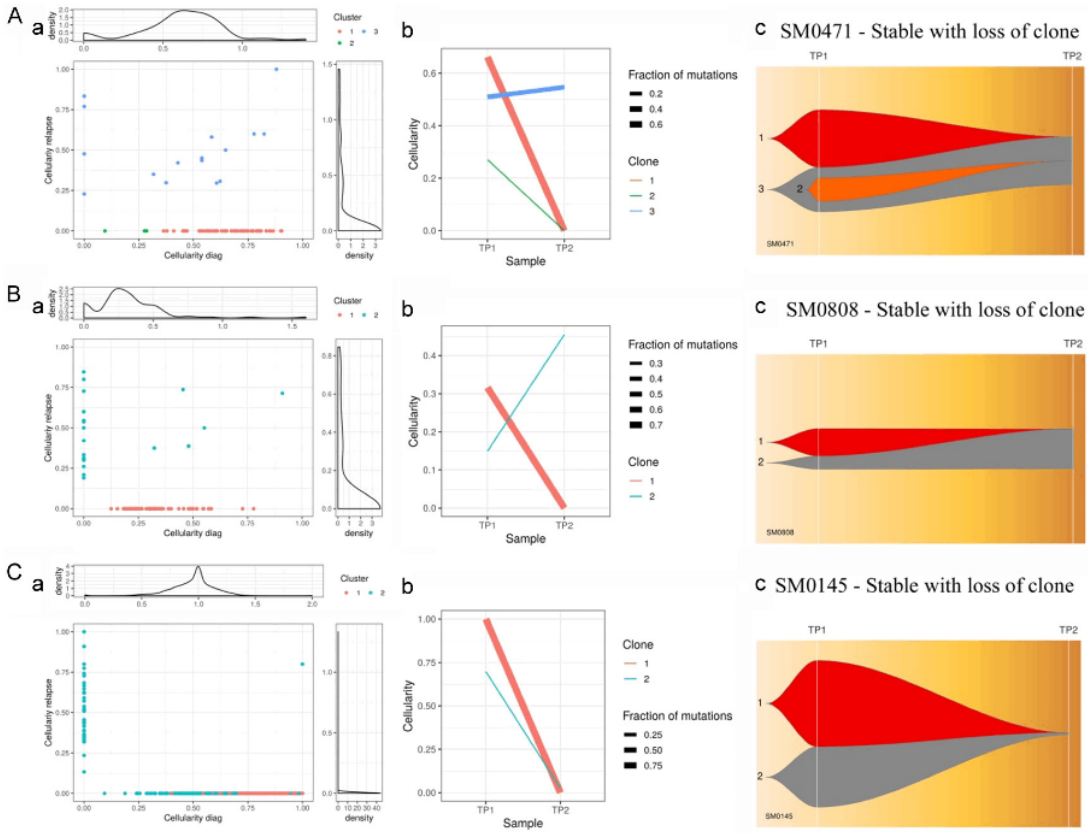


Figure B.13: (A-E) Clonal evolution in each case of MM. Representation of clonal evolution through (a) Density, (b) Evolution and (c) Fish plots across individual MM patients with branching, linear and stable with loss of clone patterns of clonal evolution.

Appendix C

Inference of clonal trajectory in single-cell data

Robust Graph Learning

Consider a noisy single-cell data matrix, X^N, of size $m \times n$. m is the number of cells and n is the number of mutation sites. This matrix is binary in nature where '1' represents the presence of mutation and '0' denotes the absence of mutation. Our goal is to extract a denoised matrix, X^D from this noisy matrix along with E which is the error observed in matrix such that $X^N = X^D + E$. This task of denoising a noisy matrix is very well performed by applying robust PCA on the noisy matrix as done in [225]. Low rank constraints are added on the denoised matrix, X^D, because the original genotype matrix is a low rank matrix where the tumor cells are clustered together into various subclones such that there is little to no variation in the genotype of the cells within the same subclone. Further, we consider that error component, E, is sparse. However, recently an improved version has been proposed to this problem i.e. Robust Graph learning [227], where along with the denoising, an adjacency matrix is simultaneously learned. Our objective now is to recover a denoised matrix as well as to learn an adjacency graph during the denoising process. Consider L and S to be the Laplacian and similarity graph learned during the denoising step. Robust graph learning is formulated in equation C.1.

$$\begin{aligned} \min_{X^N, X^D, E} & ||X^D||_* + a||E||_1 + b\text{Tr}X^D L(X^D)^T + c||S||_F^2, \\ \text{s.t.} \quad & X^N = X^D + E, S1 = 1, 0 \leq S \leq 1 \end{aligned} \tag{C.1}$$

where a, b and c are the trade-off parameters. In the above formulation, denoising and graph learning are implemented together such that each of the step iteratively enhances the other step. In order to solve equation C.1, an auxiliary variable, W is introduced such that the equation becomes:

$$\begin{aligned} \min_{X^N, X^D, E, W} & ||X^D||_* + a||E||_1 + b\text{Tr}(WLW^T) + c||S||_F^2, \\ \text{s.t.} \quad & X^N = X^D + E, S1 = 1, 0 \leq S \leq 1, W = X^D \end{aligned} \tag{C.2}$$

The above formulation can now be solved via alternating direction method of multipliers (ADMM).

Proposed Extension of Robust Graph Learning for Recovering Missing Values

In this work, we have extended robust graph learning for handling missing entries in the data. Single-cell data not only suffers from noisy corruptions but it also has missing values which needs to be tackled for accurate downstream analysis. Therefore, we have extended the robust graph learning algorithm to handle missing values. A linear operator, $P_\Omega(X^D)$, was defined which sets the unobserved entries to 0 while keeping the rest equal to the observed entries as follows:

$$P_\Omega(X^D) = \begin{cases} X^D_{ij} & \text{if(i,j)} \in \Omega; \\ 0 & \text{if(i,j)} \notin \Omega. \end{cases} \tag{C.3}$$

Equation C.1 was modified such that the denoising and missing value imputation occurs simultaneously along with graph learning.

$$\begin{aligned} \underset{X^D,S,E}{min} & ||X^D||_* + a||E||_1 + b\text{Tr}X^D L(X^D)^T + c||S||_F^2, \\ \text{s.t.} \quad & P_\Omega(X^N) = P_\Omega(X^D + E), S1 = 1, 0 \leq S \leq 1 \end{aligned} \tag{C.4}$$

The above formulation ensures that the low rank denoised matrix and noisy sparse component is recovered from the observed entries, $P_\Omega(X^N)$. The above equation can be further converted to the following framework:

$$\begin{aligned} \underset{X^D,S,E}{min} & ||X^D||_* + a||P_\Omega(E)||_1 + b\text{Tr}X^D L(X^D)^T + c||S||_F^2, \\ \text{s.t.} \quad & X^N = X^D + E, S1 = 1, 0 \leq S \leq 1 \end{aligned} \tag{C.5}$$

The above formulation can now be solved via alternating direction method of multipliers (ADMM) by adding an auxiliary variable W in the equation.

$$\begin{aligned} \underset{X^D,S,E,W}{min} & ||X^D||_* + a||P_\Omega(E)||_1 + b\text{Tr}(WLW^T) + c||S||_F^2, \\ \text{s.t.} \quad & X^N = X^D + E, S1 = 1, 0 \leq S \leq 1, W = X^D \end{aligned} \tag{C.6}$$

Augmented Lagrangian function can be obtained by removing equality constraints on X^N and W:

$$\mathcal{L}(X^D, E, S, W, Z_1, Z_2) = ||X^D||_* + a||P_\Omega(E)||_1 + b\text{Tr}(WLW^T) + c||S||_F^2 + \frac{\mu}{2}\left(||X^D + E - X^N + \frac{Z_1}{\mu}||_F^2 + ||X^D - W + \frac{Z_2}{\mu}||_F^2\right) \quad \text{s.t.} \quad S1 = 1, 0 \leq S \leq 1 \tag{C.7}$$

where μ is penalty parameter and Z_1 and Z_2 are the Lagrangian multipliers. In our proposed method, ARCANE-ROG, a was set to $(1 + 3 \times \Omega)/\sqrt{m \times n}$, b was set to $5/\sqrt{m \times n}$ and c was set to $5/\sqrt{m \times n}$. The above function can be solved iteratively for each of the parameters one by one by keeping the other parameters fixed as follows.

- **Update** X_D: We update X^D after fixing other variables such that problem (C.7) becomes

$$\underset{X^D}{min}||X^D||_* + \mu||X^D - X^I||_F^2 \tag{C.8}$$

where X^I is X^I = [($X^N + W - E -$ ($Z_1 + Z_2$) $/\mu$) $/2$] . It has a closed form solution according to singular value shrinkage, i.e.

$$X^D = U\text{diag}((\sigma - (\frac{1}{2\mu}))_+)V^T \tag{C.9}$$

where

$$U\text{diag}(\sigma)V^T \text{ is SVD of } \quad X^I = \left(X^N + W - E - \frac{(Z_1 + Z_2)}{\mu}\right)/2 \tag{C.10}$$

- **Update** E: We update E after fixing other variables such that problem (C.7) becomes

$$\underset{E}{mina}||P_\Omega(E)||_1 + \frac{\mu}{2}||E - \left(X^N - X^D - \frac{Z_1}{\mu}\right)||_F^2 \tag{C.11}$$

It also admits closed-form solution, i.e.

$$\begin{aligned} e_{ij} &= (|o_{ij}| - \frac{a}{\mu})_+.sign(o_{ij}) \\ \text{where } O &= X^N - X^D - \frac{Z_1}{\mu} \end{aligned} \tag{C.12}$$

It is to be noted that sparse component, E, were recovered only from the observations i.e. $P_\Omega(X^D)$.

- **Update** S: L is also a function of S, so for updating S, equation (C.7) becomes

$$\underset{s_i}{min} \sum_{j=1}^{n} \left(\frac{b}{2}||w_i - w_j||^2 s_{ij} + c s_{ij}^2\right) \text{ s.t. } s_i^T 1 = 1, 0 \leq s_{ij} \leq 1. \tag{C.13}$$

We can denote $||w_i - w_j||^2$ as f_{ij} where $f_i \in \mathcal{R}^{n \times 1}$. Thus, the above can be

reformulated as

$$\underset{s_i^T 1=1, 0\leq s_{ij}\leq 1}{min} ||s_i - \frac{b}{4c} f_i||^2 \tag{C.14}$$

The problem has a sparse solution which is why we update the k-nearest neighbours, that it, s_j has k positive entries and $s_{ij} = 0$ for $j > k$. The Lagrangian function of it is

$$\mathcal{L}(s_i, \eta, \xi) = ||s_i + \frac{b}{4c} f_i||^2 - \eta(s_i^T 1 - 1) - \xi^T s_i \tag{C.15}$$

where η and $\xi \in \mathcal{R}^{n\times 1}$ are the Lagrangian multipliers and the overall c can be set to the average of $\{c_i\}_{i=1}^n$. By the Karush-Kuhn-Tucker condition, it yields $s_i = ((\eta/2) - (bf_i/4c_i))_+$. We then rank f_i in ascending order and we obtain

$$\begin{cases} s_{ik} = \frac{\eta}{2} - \frac{bf_{ik}}{4c_i} > 0 \\ s_{i,k+1} = \frac{\eta}{2} - \frac{bf_{i,k+1}}{4c_i} \leq 0 \\ s_i^T 1 = \sum_{j=1}^k (\frac{\eta}{2} - \frac{bf_{ij}}{4c_i}) = 1 \end{cases} \tag{C.16}$$

$$\Rightarrow \begin{cases} s_{ij} = \frac{f_{i,k+1} - f_{ij}}{k f_{i,k+1} - \sum_{r=1}^k f_{ir}}, j \leq k \\ c_i = \frac{b}{4}\Big(k f_{i,k+1} - \sum_{j=1}^k f_{ij}\Big) \\ \eta = \frac{2}{k} + \frac{b}{2kc_i} \sum_{j=1}^k f_{ij} \end{cases} \tag{C.17}$$

The value of c_i has been set to maximum in the above derivation. Thus, taking the average of $\{c_i\}_{i=1}^n$, we have

$$c = \frac{b}{4n} \sum_{i=1}^n \Big(k f_{i,k+1} - \sum_{j=1}^k f_{ij}\Big) \tag{C.18}$$

- **Update** Z: For updating Z we have

$$\underset{W}{min} b\text{Tr}(WLW^T) + \frac{\mu}{2} ||E - \Big(X^D - W - \frac{Z_2}{\mu}\Big)||_F^2 \tag{C.19}$$

Its first-order derivative is $2bWL - \mu(X^D - W + (Z_2/\mu))$. By setting it to zero, we achieve

$$W = (\mu X^D + Z_2)(2bL + \mu I)^{-1} \tag{C.20}$$

- **Update Lagrangian multipliers**:

$$\begin{aligned} Z_1 &= Z_1 + \mu(X^D + E - X^N) \\ Z_2 &= Z_2 + \mu(X^D - W) \end{aligned} \tag{C.21}$$

Algorithm 2 Algorithm for denoising and imputation of noisy and incomplete data

Input: X^N (Incomplete and noisy matrix), parameters $a > 0$, $b > 0$, $\mu > 0$, c
Initialize: $W = X^N$, $E = 0$, $Z_1 = Z_2 = 0$
while converge **do**
  Calculate X^D as:
    $X^D = U\text{diag}((\sigma - (\frac{1}{2\mu}))_+)V^T$,
      where $U\text{diag}(\sigma)V^T$ is SVD of $A = (X^N + W - E - \frac{(Z_1+Z_2)}{\mu})/2$
  Update E as:
    $e_{ij} = (|o_{ij}| - \frac{a}{\mu})_+.sign(o_{ij})$, where $O = X^N - X^D - \frac{Z_1}{\mu}$
  Update S as:
    $s_{ik} = \frac{f_{i,k+1} - f_{ij}}{k f_{i,k+1} - \sum_{r=1}^{k} f_{ir}}, j \leq k$ where $f_{ij} = ||w_i - w_j||^2$
  Update W as:
    $W = (\mu X^D + Z_2)(2bL + \mu I)^{-1}$
  Update Lagrangian multipliers as:
    $Z_1 = Z_1 + \mu(X^D + E - X^N)$
    $Z_2 = Z_2 + \mu(X^D - W)$
end while
Output: X^D (Denoised data), S (Similarity matrix/ Adjacency graph)

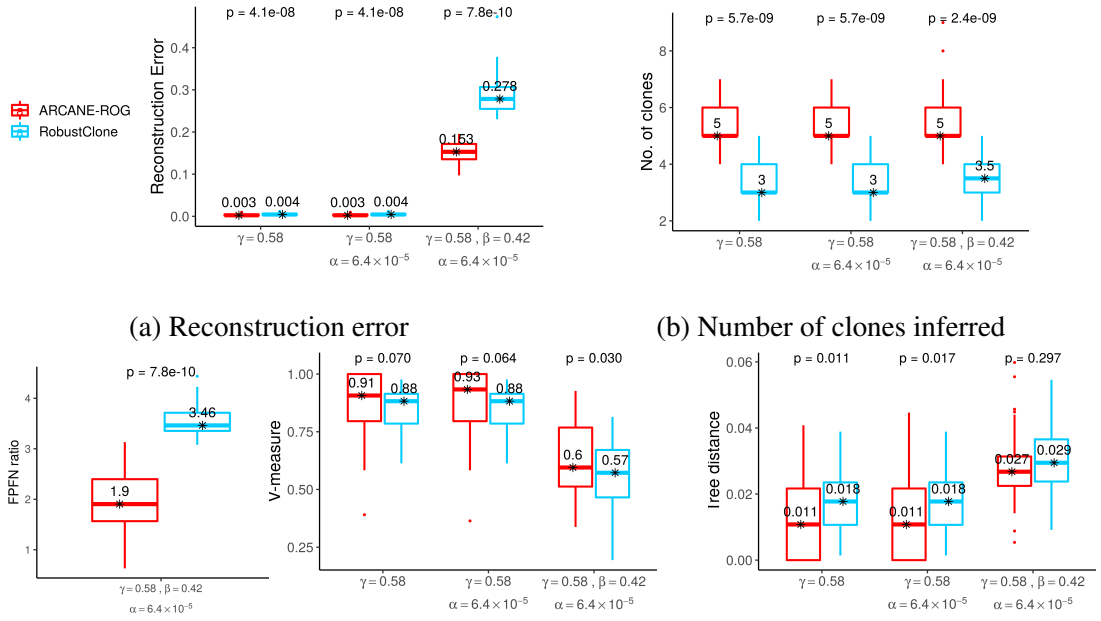
Additional results

Results on simulated datasets for real datasets

There is no ground truth in real single cell datasets. Therefore, to test the performance of our proposed method on real datasets, we simulated data imitating the characteristics of real datasets in terms of the missing values, false positives and false negatives. Size of the simulated datasets was fixed to the size of the real datasets. We applied RobustClone and ARCANE-ROG on the simulated datasets and found that ARCANE-ROG performed significantly (p-value < 0.5) superior to RobustClone for all the datasets under different conditions. Overall comparison of the both the methods is shown in the Table 6.2.

Results on Real datasets

Five clones were inferred in clear-cell renal-cell carcinoma dataset. The sequence in which the mutations were acquired are shown in the Figure C.7. FGFR4 gene in red denotes an actionable mutation and is visible in the initial stages of the mutations tree in the data in Figures C.7(A) and C.7(B). However, mutations in genes like NOS1 and CDON are shown in the bottom of the hierarchy in the figure C.7(B) suggesting that they are acquired in the later stages of cancer compared to what is observed in the fig-



(a) Reconstruction error (b) Number of clones inferred

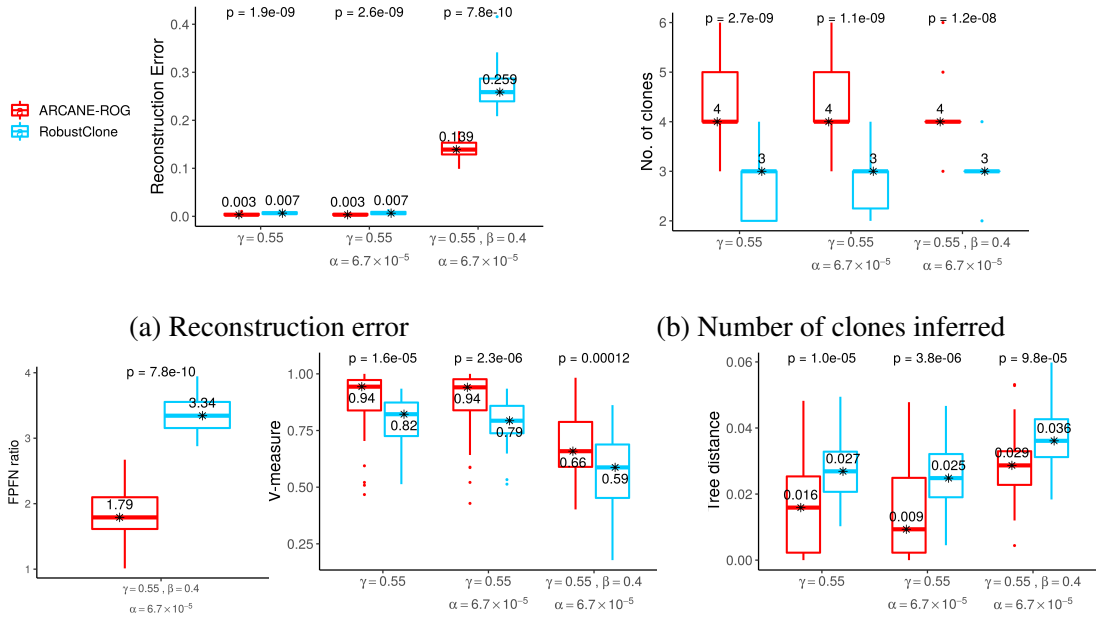
(c) FPFN ratio (d) V-measure (e) Tree distance

Figure C.1: Boxplots for comparison of the proposed ARCANE-ROG method with RobustClone for simulated dataset generated for JAK2-negative myeloproliferative neoplasm data of size 58×712. (a) Reconstruction error when the data has the maximum noise as it was being corrupted with missing values and false positives as well false negatives. (b) FPFN ratio was calculated only when the simulated data had both the false positives and false negatives along with the missing entries. (c) Number of clones estimated by ARCANE-ROG were close to the actual number of clones while RobustClone underestimated the number of clones. (d) V-measure was the lowest when the added noise was the highest (e) Tree distance gradually increased with an increase in noise. Overall, the performance of ARCANE-ROG was significantly (p-value < 0.5) better than RobustClone.

ure C.7(A). Similarly, genes like PTPRT which are altered in later stages in C.7(A) are found to altered in the initial stages in C.7(B). As we have already discussed in the main manuscript that there is limited knowledge on predetermined order of mutations, but, the sequential acquisition of the mutations does influence cancer progression. Therefore, inferring an accurate sequence of mutation may assist us in drawing more relevant and significant biological findings from the data.

Comparison with BnpC algorithm

We have also compared our method with BnpC [228] algorithm. However, the method was computationally expensive even on small datasets. Therefore, we ran BnpC only on single dataset of each simulated case. BnpC took nearly 12 hrs on data of size 500×500 while ARCANE-ROG takes less than 2 min for the same task. Default settings were



(a) Reconstruction error

(b) Number of clones inferred

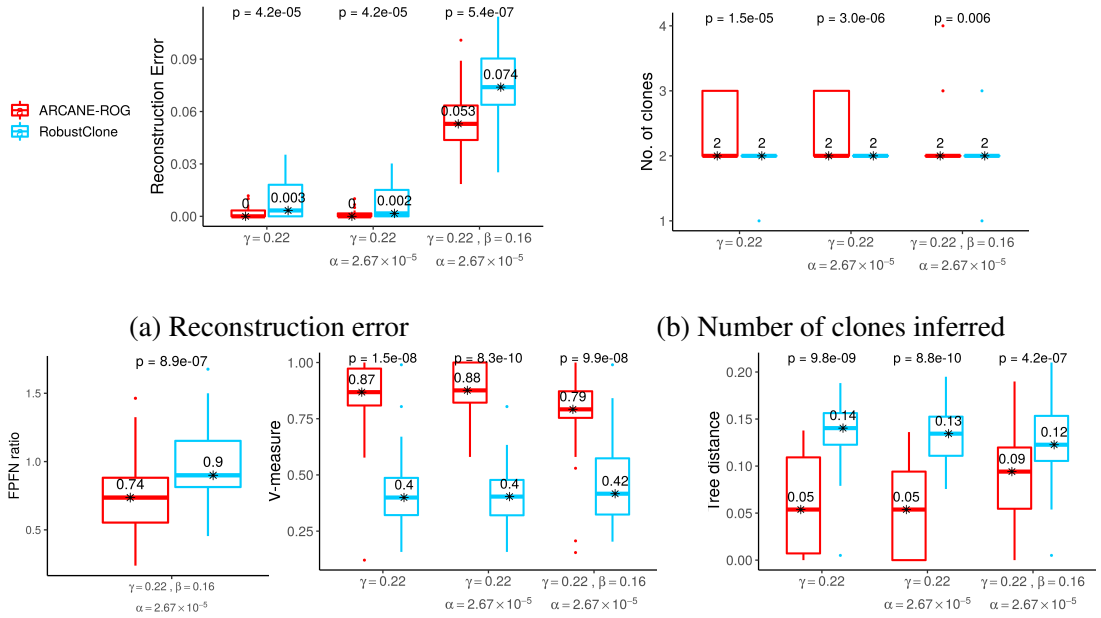
(c) FPFN ratio

(d) V-measure

(e) Tree distance

Figure C.2: Boxplots for comparison of the proposed ARCANE-ROG method with RobustClone for simulated dataset generated for Muscle-invasive bladder transitional cell carcinoma data of size 44×443. (a) Reconstruction error when the data has the maximum noise as it was being corrupted with missing values and false positives as well false negatives. (b) FPFN ratio was calculated only when the simulated data had both the false positives and false negatives along with the missing entries. (c) Number of clones estimated by ARCANE-ROG were close to the actual number of clones while RobustClone underestimated the number of clones. (d) V-measure was the lowest when the added noise was the highest (e) Tree distance gradually increased with an increase in noise. Overall, the performance of ARCANE-ROG was significantly (p-value < 0.5) better than RobustClone.

used to run BnpC and 20 cores were alloted for the process in an Ubuntu system with 98 GB RAM. We have compared BnpC with ARCANE-ROG in terms of tree distance error and V-measure.



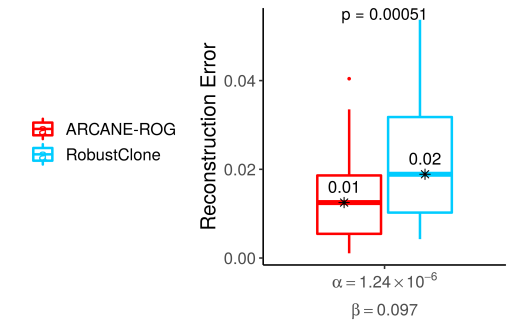
(a) Reconstruction error (b) Number of clones inferred

(c) FPFN ratio (d) V-measure (e) Tree distance

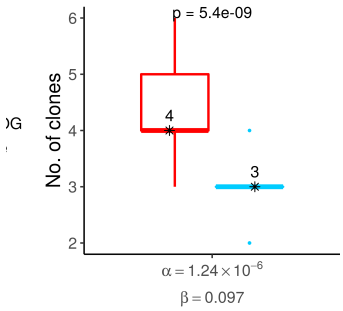
Figure C.3: Boxplots for comparison of the proposed ARCANE-ROG method with RobustClone for simulated dataset generated for real Clear-cell renal-cell carcinoma data of size 17×35. (a) Reconstruction error when the data has the maximum noise as it was being corrupted with missing values and false positives as well false negatives. (b) FPFN ratio was calculated only when the simulated data had both the false positives and false negatives along with the missing entries. (c) Number of clones estimated by ARCANE-ROG were close to the actual number of clones while RobustClone underestimated the number of clones. (d) V-measure was the lowest when the added noise was the highest (e) Tree distance gradually increased with an increase in noise. Overall, the performance of ARCANE-ROG was significantly (p-value < 0.5) better than RobustClone.

Performance of BnpC on varying α, β, γ, number of mutation sites, number of cells and clones.

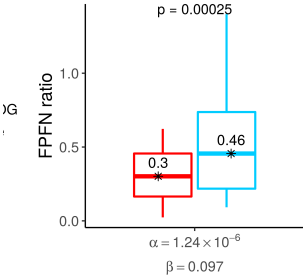
With increase in α, tree distance increased and V-measure decreased as shown in the table C.1. Overall, ARCANE-ROG performed superior to BnpC in terms of tree distance error and V-measure. Similarly at all values of β, BnpC has high tree distance and low V-measure as compared to ARCANE-ROG. For different values of γ, ARCANE-ROG performs better than BnpC as it has low tree distance and higher value of V-measure at all values of γ. When mutation sites are equal to 100, BnpC performs better as compared to ARCANE-ROG, however, for higher number of mutations sites, our proposed method has low tree distance error and high V-measure. With change in number of cells and clones, BnpC performs slightly better as compared to ARCANE-ROG, however, the computational time 12 hrs which is very high in comparison to ARCANE-ROG which takes less than 5 minutes for the computation. So, it is evident from the above



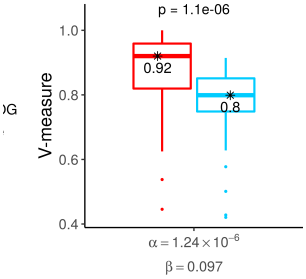
(a) Reconstruction error



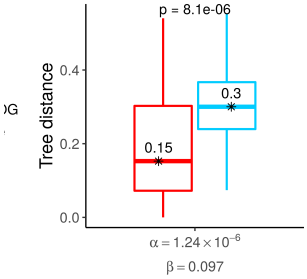
(b) Number of clones inferred



(c) FPFN ratio

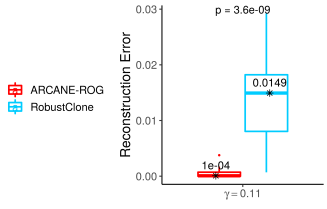


(d) V-measure

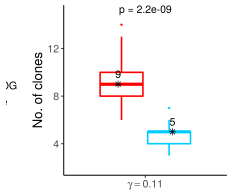


(e) Tree distance

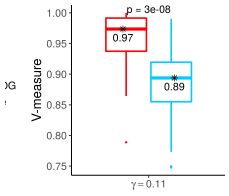
Figure C.4: Boxplots for comparison of the proposed ARCANE-ROG method with RobustClone for simulated dataset generated for real dataset, 47×40. ARCANE-ROG exhibited superior performance as compared to RobustClone in terms of low values of (a) Reconstruction error, (b) FPFN ratio, and (e) Tree distance. (c) Number of clones inferred by ARCANE-ROG were close to the actual number of clones while RobustClone underestimated the number of clones. (d) V-measure was higher for our proposed method. Overall, the performance of ARCANE-ROG was significantly (p-value < 0.5) better than RobustClone.



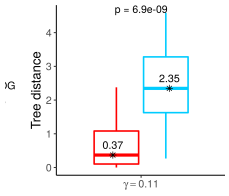
(a) Reconstruction error



(b) Number of clones inferred

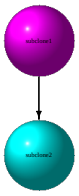


(c) V-measure

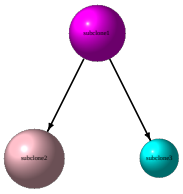


(d) Tree distance

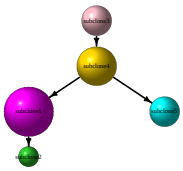
Figure C.5: Boxplots for comparison of the proposed ARCANE-ROG method with RobustClone for simulated dataset generated for High grade serious ovarian cancer dataset with size 420×48. (a) Reconstruction error and (d) Tree distance values are low for our proposed method. (b) Number of clones estimated by ARCANE-ROG were close to the actual number of clones while RobustClone underestimated the number of clones (c) V-measure for ARCANE-ROG is high. Overall, the performance of ARCANE-ROG was significantly (p-value < 0.5) better than RobustClone.



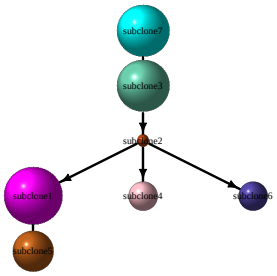
(a) JAK2-negative myeloproliferative neoplasm (58 cells and 712 mutations)



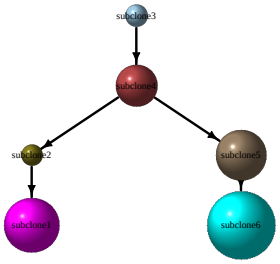
(b) Muscle-invasive bladder transitional cell carcinoma (44 cells and 443 mutations)



(c) Clear-cell renal-cell carcinoma (17 cells and 35 mutations)



(d) Estrogen-receptor positive (ER+) breast cancer (47 cells and 40 mutations)



(e) High grade serious ovarian cancer (420 cells and 48 mutations)

Figure C.6: Performance of ARCANE-ROG on real datasets. The subclones inferred in the data and the pattern of clonal trajectory inferred via ARCANE-ROG.

experiments that ARCANE-ROG is robust to varying α, β, γ, number of mutation sites, number of cells and clones. Further, it is computationally robust and efficient in terms of performance as well as time.

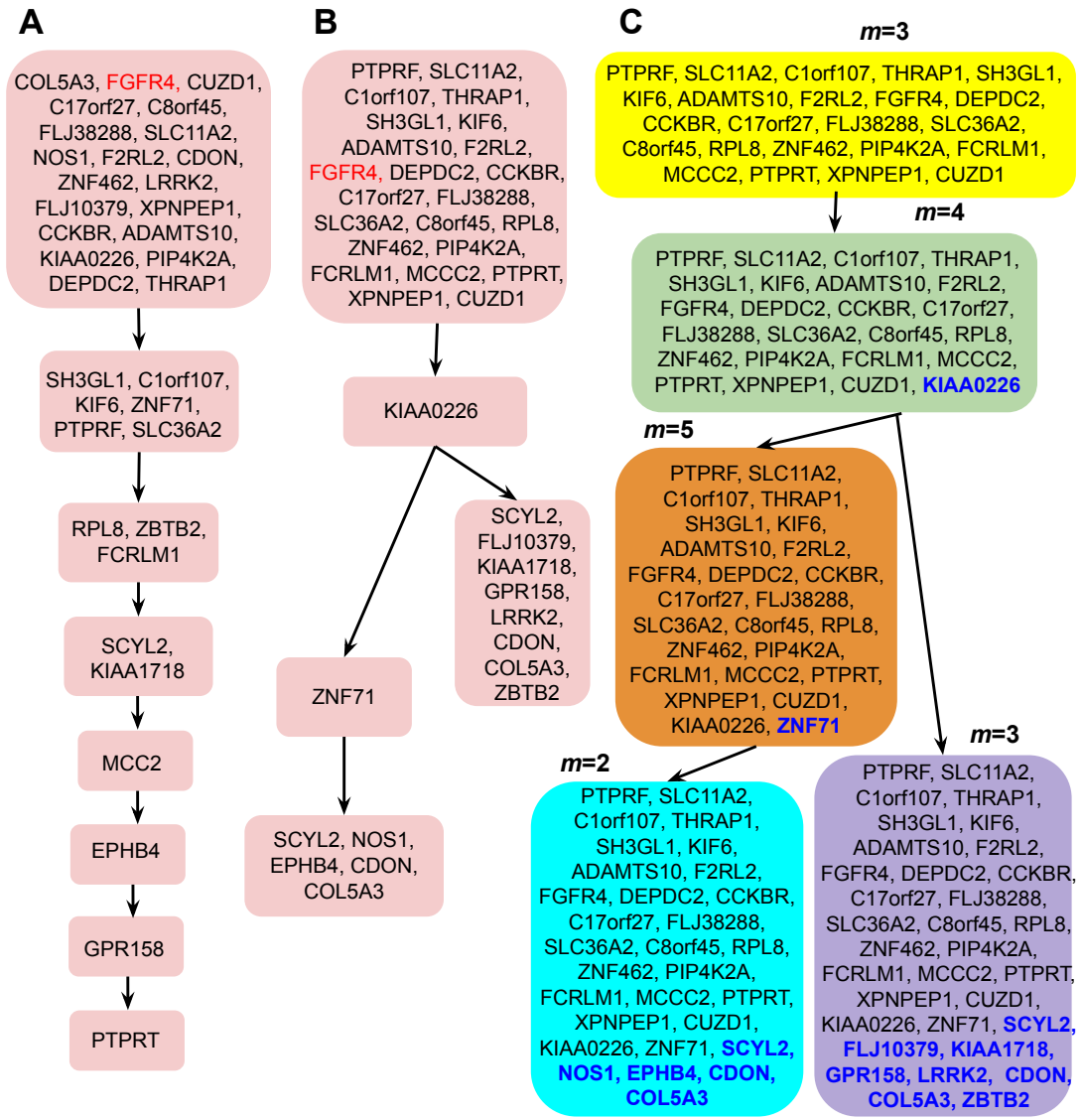


Figure C.7: Comparison of the results obtained on real dataset of clear-cell renal-cell carcinoma. A. Maximum Likelihood (ML) tree for the clear-cell renal-cell carcinoma dataset. Mutations placed in a single box have non-identifiable order. B. Sequence of mutations inferred via ARCANE-ROG. Red genes indicate actionable mutations according to TARGET/COSMIC database C. Clonal tree deduced via ARCANE-ROG. Five clones were inferred in total. Child clone has all the mutations acquired in the parent clone. New mutations acquired by the child clone are shown in the blue color. *m* denotes the number of cells.

Table C.1: Performance comparison between BnpC and ARCANE-ROG for different values of α, β, γ, m, n and s. ARCANE-ROG is more robust to BnpC with low tree distance error and high V-measure at all settings.

		Tree distance		V-measure	
		BnpC	ARCANE-ROG	BnpC	ARCANE-ROG
α	0.001	0	0	1	1
	0.01	0.013832	0	0.986	1
	0.1	0.0522	0.002992	0.968	0.99
β	0.1	0.278424	0	0.822	1
	0.2	0	0	1	1
	0.3	0	0	1	1
	0.4	0.06846	0.015936	0.942	0.978
γ	0.2	0	0	1	1
	0.3	0.081992	0.004	0.948	0.99
	0.4	0.007984	0	0.99	1
	0.5	0	0	1	1
n	100	0.04972	0.24752	0.979	0.90
	500	0	0	1	1
	1000	0.023612	0	0.96	1
m,s	100 cells, 10 clones	0.01016	0.02116	0.98	0.94
	500 cells, 20 clones	0	0.017896	1	0.989
	1000 cells, 30 clones	0.01394	0.12288	0.998	0.966
	2000 cells, 40 clones	0.66056	0.504918	0.988	0.96