ABSTRACT

Title of Dissertation: DISSECTING TUMOR CLONALITY IN LIVER CANCER: A PHYLOGENY ANALYSIS USING COMPUTATIONAL AND STATISTICAL TOOLS

Zeynep Kacar
Doctor of Philosophy, 2023

Dissertation Directed by: Professor Eric Slud
Department of Mathematics and Statistics

Liver cancer is a heterogeneous disease characterized by extensive genetic and clonal diversity. Understanding the clonal evolution of liver tumors is crucial for developing effective treatment strategies. This dissertation aims to dissect the tumor clonality in liver cancer using computational and statistical tools, with a focus on phylogenetic analysis. Through advancements in defining and assessing phylogenetic clusters, we gain a deeper understanding of the survival disparities and clonal evolution within liver tumors, which can inform the development of tailored treatment strategies and improve patient outcomes.

The thesis begins by providing an overview of sources of heterogeneity in liver cancer and data types, from Whole-Exome (WEX) and RNA sequencing (RNA-seq) read-counts by gene to derived quantities such as Copy Number Alterations (CNAs) and Single Nucleotide Variants (SNVs). Various tools for deriving copy-numbers are discussed and compared. Additionally, comparison of survival distributions is discussed.

The central data analyses of the thesis concern the derivation of distinct clones and clustered phylogeny types from the basic genomic data in three independent cancer cohorts, TCGA-LIHC, TIGER-LC and NCI-MONGOLIA. The SMASH (Subclone multiplicity allocation and somatic heterogeneity) algorithm is introduced for clonality analysis, followed by a discussion on clustering analysis of nonlinear tumor evolution trees and the construction of phylogenetic trees for liver cancer cohorts. Identification of drivers of tumor evolution, and the immune cell micro-environment of tumors are also explored.

In this research, we employ survival analysis tools to investigate and document survival differences between groups of subjects defined from phylogenetic clusters. Specifically, we introduce the log-rank test and its modifications for generic right-censored survival data, which we then apply to survival follow-up data for the subjects in the studied cohorts, clustered based on their genomic data. The final chapter of this thesis takes a significant step forward by extending an existing methodology for covariate-adjustment in the two-sample log-rank test to a K-sample scenario, with a specific focus on the already defined phylogeny cluster groups. This extension is not straightforward because the computation of the test statistic for K-sample and its asymptotic null distribution do not follow directly from the two-sample case. Using these extended tools, we conduct an illustrative data analysis with real data from the TIGER-LC cohort, which comprises subjects with analyzed and clustered genomic data, leading to defined phylogenetic clusters associated with two different types of liver cancer. By applying the extended methodology to this dataset, we aim to effectively assess and validate the survival curves of the defined clusters.

DISSECTING TUMOR CLONALITY IN LIVER CANCER: A PHYLOGENY ANALYSIS USING COMPUTATIONAL AND STATISTICAL TOOLS

by

Zeynep Kacar

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Dr. Eric Slud, Chair/Advisor
Dr. Doron Levy, Co-Advisor
Dr. Xin Wei Wang
Dr. Joan Jian-Jian Ren
Dr. Najib M. El-Sayed, Dean's Representative

© Copyright by
Zeynep Kacar
2023

Acknowledgments

I am deeply grateful to all the individuals who have played a vital role in making this thesis possible. First and foremost, I would like to express my sincere appreciation to my advisor, Professor Eric Slud. His unwavering support and belief in me have been instrumental throughout this research. Professor Slud's availability, guidance, and invaluable advice have been constant sources of inspiration. It has been an honor and privilege to collaborate with such an exceptional mentor and scholar. I would also like to extend my gratitude to my co-advisors, Dr. Doron Levy and Dr. Xin Wang, for their guidance and expertise, which have greatly enriched the quality and depth of this work. I am thankful to my committee members, Dr. Ren and Dr. El-seyad, for their time and effort in reviewing my thesis. My heartfelt thanks go to my family, especially my mother and father. Their unending support and guidance have been instrumental in shaping my career and have provided me strength during challenging times. I am forever indebted to them, and words cannot adequately express my gratitude. Lastly, I would like to express my deepest appreciation to my soulmate, Ergun Kacar. Your presence and support have been a constant source of strength and encouragement throughout this journey. I am also immensely grateful to my angels, Zehra and Sait Ali, whose boundless love and energy have propelled me to overcome every obstacle. To all those who have contributed in ways seen and unseen, thank you for being part of this incredible journey. Your support and belief in me have been the foundation of my success, and I am forever grateful.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

WEX	Whole Exome Sequencing	1
RNA-seq	RNA Sequencing	1
CNA	Copy Number Alterarion	1
SNV	Single Nucleotide Variant	1
SMASH	Subclone Multiplicity Allocation and Somatic Heterogeneity	2
SNP	Single Nucleotide Polymorphism	1
BAM	Binary Alignment Map	8
BWA	Burrows-Wheeler Aligner	8
GATK	Genome Analysis Toolkit	8
VCF	Variant Call Format	8
cDNA	Complementary DNA	12
mRNA	Messenger RNA	14
aCGH	Array Comparative Genomic Hybridization	15
HCC	Hepatocellular Carcinoma	15
CCA	Cholangiocarcinoma	15
GDC	Genomic Data Commons Data Portal portal	17
AGNES	Agglomerative Nesting	22
DIANA	Divisive Analysis	22
PAM	Partitioning Around Medoids	24
PCF	Piecewise Constant Fitting	34
CBS	Circular Binary Segmentation	34
HMM	Hidden Markov Model	34
ASCAT	Allele-Specific Copy number Analysis of Tumors	33
FACETS	Fraction and Allele-Specific Copy Number Estimates from Tumor Sequencing	33
PCA	Principal Component Analysis	78
PE	Proportion Entropy	70
ME	Mutation Entropy	75
ITH	Intra-Tumor Heterogeneity	92
TME	Tumor Micro-environment	65

Chapter 1: Introduction

1.1 Overview

The clonal evolution and heterogeneity of tumors have emerged as key factors impacting cancer progression, treatment response, and patient prognosis (McGranahan and Swanton, 2017). With advancements in next-generation sequencing technologies, the ability to analyze tumor genomes at high resolution has significantly expanded. These technological breakthroughs have facilitated the identification and characterization of clonal populations within liver cancer, enabling a deeper understanding of tumor development, evolution, and metastasis.

Liver cancer is a significant global health concern, with its incidence and mortality rates steadily increasing over the years (Sung et al., 2021). In-depth research focusing on genomic data analysis and exploration of survival patterns is crucial to better understand the underlying mechanisms of tumor progression and improve patient outcomes. The genomic data analyzed in this research consists of high-dimensional datasets that capture genetic information at various levels. These datasets include data from WEX (Whole Exome Sequencing), RNA-seq (RNA sequencing), as well as SNP-arrays (Single Nucleotide Polymorphisms). The details of these data sets are provided in Sections 1.3, 1.4, and 1.5.

Our analysis incorporates clinical information, demographic data, and other pertinent metadata linked to the individuals or samples being studied. This comprehensive approach involves processing

and integrating these diverse types of genomic data to obtain a comprehensive understanding of the genetic landscape and its intricate connection with disease-related outcomes.

This thesis aims to explore the complex and interconnected nature of tumor cells by employing an approach called phylogenetic tree representation. In simpler terms, we will investigate the genetic lineage of tumor cells based on the specific mutations they accumulate. These mutations serve as markers that help us trace the paths of genetic evolution within the tumor. A phylogenetic tree reveals the evolutionary relationships among unique group of tumor cells (”clones”) within a patient. Through clonality and clustering analysis, the study identifies three distinct groups within the set of phylogenetic trees. The objective is to provide evidence for assessing and justifying the validity of these defined groups through the application of techniques of statistical hypothesis testing. By employing these statistical techniques, the study aims to validate the differences observed among the phylogenetic tree groups. The research contributions of the submitted work of this thesis are structured along the defined pipeline in Figure 1.1, encompassing several key components.

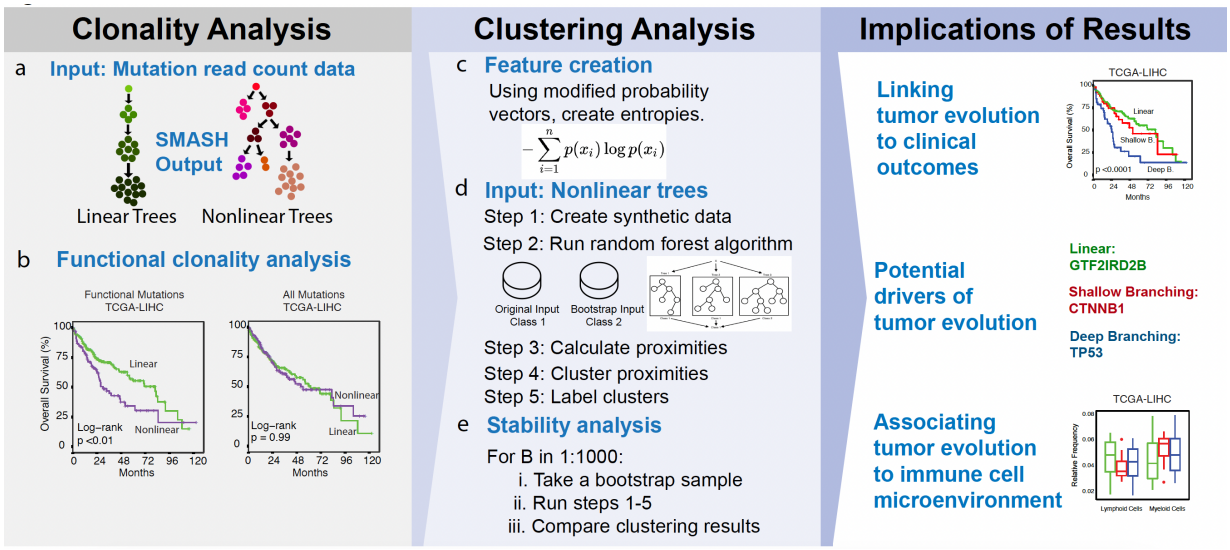


Figure 1.1: Graphical Summary of Clonality and Clustering Analysis for Liver Cancer Cohorts

First, the thesis presents a clonality analysis of liver cancer using a particular set of sequencing data from three liver cancer cohorts collected under the auspices of Dr. Xin Wang's Liver Cancer Laboratory at the National Cancer Institute. To ensure accurate and reliable results, data preprocessing is performed, specifically restricting the analysis to functional mutations. Moreover, Copy Number Alteration (CNA) analysis is prioritized, as it plays a significant role in quantifying the presence and abundance of mutated genes. An existing software, Sequenza, (Favero et al., 2015) for CNA analysis is utilized in this work. The thesis also provides an overview and comparison of popular CNA tools on liver cancer data. The clonality analysis investigates the underlying clonal architecture and evolutionary dynamics of liver tumors, contributing to a valuable understanding of clonal diversity, the mutational landscape, and potential driver events involved in the progression of liver cancer.

Second, the research focuses on the definition of new features in preparation for clustering the linear and nonlinear phylogenetic tree groups obtained from the clonality analysis. Tumor evolution has long been understood primarily through the lens of the linear evolution model, as proposed by (Nowell, 1976). According to this model, tumors accumulate clonal mutations with highly dominant selective properties, resulting in the out-competing of all previous clones. This linear accumulation of clonal mutations was considered the prevailing paradigm in tumor evolution for a significant period. However, observations from several studies (Dexter et al., 1978; Heppner, 1984) challenged the assumption of strict linearity in tumor growth. These studies demonstrated that tumors can exhibit nonlinear growth patterns, characterized by the presence of multiple molecularly distinct subclones. Moreover, the classification of tumor evolution models extends beyond a simple binary categorization. Recent research studies, (Davis et al., 2017; Vendramin et al., 2021; Zhu et al., 2021), have contributed to a more comprehensive understanding

of tumor evolution by refining the classification and incorporating various distinct models. These models encompass a range of evolutionary patterns, including linear evolution, branching evolution, neutral evolution, and macroevolution. To enable effective clustering of nonlinear phylogenetic tree groups, new features are defined in this thesis. These features are designed to capture the specific characteristics and complexities associated with the different types of tumor evolution. We utilize Random Forests (Breiman, 2001) as a way of clustering, which is explained in Section 3.3.2, to identify meaningful groups within the nonlinear phylogenetic trees. We refer to the Random Forest authors' website (Breiman and Cutler, 2023) where an explanation is provided on how to implement Random Forests effectively for clustering purposes. To assess the separation of the identified clonal tree clusters, principal component analysis (PCA) and scatter plots are employed. These visualizations allow for the examination of the distribution and overlap of the clusters in a reduced-dimensional space. This analysis provides evidence for the distinct separation of the identified clonal tree clusters and strengthens the overall analysis of clonality. Furthermore, the stability of the identified clusters is assessed using the bootstrap method. This re-sampling technique involves repeatedly sampling subsets of the data to generate multiple bootstrap samples. The clustering algorithm is then applied to each bootstrap sample to evaluate the consistency and stability of the resulting clusters.

Third, our research aims to ascertain the validity and stability of our findings from cluster analysis by employing the log-rank test and its modifications as a hypothesis testing procedure. The application of survival analysis methods has become increasingly essential in evaluating the impact of various factors on patient outcomes (Collett, 1994). The log-rank test, a widely employed statistical technique, has proven valuable in assessing the survival differences between different groups. By incorporating covariate-adjusted (Kong and Slud, 1997; Ye et al., 2023) and

multi-sample stratified log-rank testing ((Fleming and Harrington, 1991)), this thesis attempts to validate and strengthen the assessment of the clonality and clustering analysis. We begin by providing a clear methodological definition of the log-rank test and its modifications. Subsequently, we explore how these statistical techniques can be effectively applied to assess the survival differences between the identified clonal clusters. The contribution of this thesis lies in presenting an approach to examining group survival differences by incorporating covariates and utilizing multi-cohort data. This approach enables the evaluation of consistent effects across different liver cancer patient cohorts, providing a better understanding of the factors influencing patient outcomes.

1.2 Sources of heterogeneity in liver cancer

Liver cancer is a complex and heterogeneous disease, characterized by variations in its characteristics among individuals and even within tumors from the same individual. This heterogeneity stems from various sources and contributes to diverse clinical outcomes and responses to treatment.

Genetic heterogeneity plays a prominent role in liver cancer, manifested by diverse genetic alterations within tumors. These alterations encompass mutations, copy number changes, chromosomal rearrangements, and epigenetic modifications. Genetic heterogeneity arises from clonal evolution, where subclones with distinct genetic changes emerge and coexist within a tumor. Liver tumors often comprise multiple clonal populations, each harboring its own unique set of genetic alterations. These clonal populations can exhibit differential growth rates, metastatic potential, and sensitivity to treatment. Clonal heterogeneity fosters intra-tumor diversity and can

profoundly impact disease progression and therapeutic outcomes.

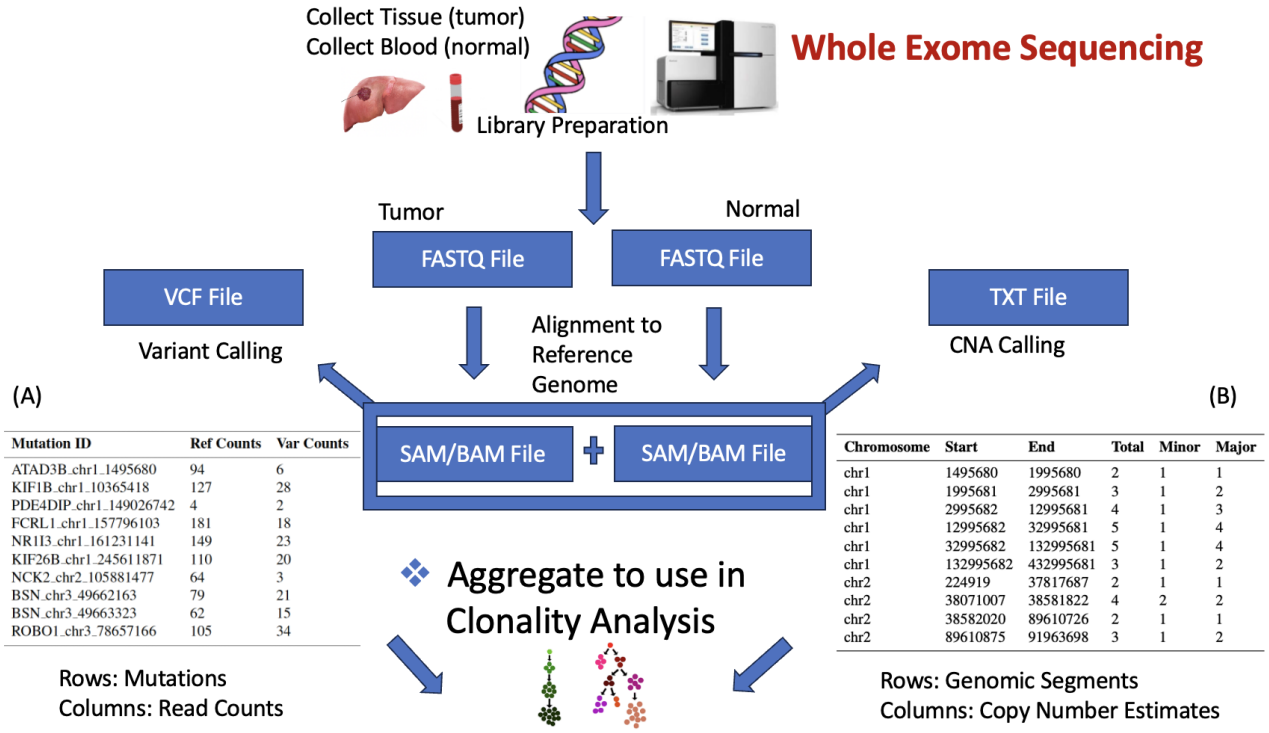
Moreover, the microenvironment surrounding liver tumors forms a complex network of cells. Heterogeneity within this microenvironment, such as variations in oxygen levels, nutrient availability, and immune cell infiltration, exerts influence on tumor behavior and response to therapy. The interplay between tumor cells and their microenvironment contributes to the observed heterogeneity.

Additionally, the heterogeneity in liver cancer arises from various etiological factors. Demographic factors like age, gender, and ethnicity, as well as clinical factors such as obesity, diet, smoking, and alcohol intake history, contribute to the heterogeneity. Furthermore, hereditary conditions like hereditary hemochromatosis and environmental factors like viral infections (such as hepatitis), liver fluke and other parasites, chemical carcinogens, and microbiota, all play a role in the vast molecular heterogeneity observed across patients. These different causative factors elicit distinct molecular mechanisms, either independently or in combination, to initiate malignant transformation, further contributing to the overall heterogeneity of liver cancer.

The extensive heterogeneity in liver cancer poses significant challenges for diagnosis, prognosis, and treatment. Understanding and characterizing this heterogeneity are crucial for developing personalized approaches to manage the disease effectively. Advances in genomic technologies and comprehensive molecular profiling are providing valuable insights into the complex landscape of liver cancer heterogeneity, paving the way for targeted therapies and precision medicine strategies tailored to individual patients.

1.3 Whole Exome Sequencing (WEX)

WEX is a powerful genomic technique that focuses on sequencing the exome, which comprises the protein-coding regions of the genome (Warr et al., 2015). The exome represents approximately 1-2% of the entire genome but it is thought to contain a vast majority of disease-causing mutations (Edelson et al., 2019). The primary objective of WEX is to identify genetic variants, including single nucleotide variants (SNVs) and small insertions/deletions (indels), within the exonic regions. By analyzing these variants, researchers can identify potential clinically relevant mutations, understand the genetic basis of diseases, and study their association with various phenotypes ((Gilissen et al., 2012)).



Mutation ID	Ref Counts	Var Counts
ATAD3B_chr1_1495680	94	6
KIF1B_chr1_10365418	127	28
PDE4DIP_chr1_149026742	4	2
FCRL1_chr1_157796103	181	18
NR1I3_chr1_161231141	149	23
KIF26B_chr1_245611871	110	20
NCK2_chr2_105881477	64	3
BSN_chr3_49662163	79	21
BSN_chr3_49663323	62	15
ROBO1_chr3_78657166	105	34

Chromosome	Start	End	Total	Minor	Major
chr1	1495680	1995680	2	1	1
chr1	1995681	2995681	3	1	2
chr1	2995682	12995681	4	1	3
chr1	12995682	32995681	5	1	4
chr1	32995682	132995681	5	1	4
chr1	132995682	432995681	3	1	2
chr2	224919	37817687	2	1	1
chr2	38071007	38581822	4	2	2
chr2	38582020	89610726	2	1	1
chr2	89610875	91963698	3	1	2

Figure 1.2: Whole Exome Sequencing Workflow

As shown in the workflow presented in Figure 1.2, the process of Whole Exome Sequencing (WEX) involves a series of steps to acquire the raw sequencing data. These steps encompass

sample collection, DNA extraction, library preparation, and sequencing. Subsequently, the sequencer processes the raw sequencing data, resulting in the generation of FASTQ files. These files contain the sequence data accompanied by their corresponding quality scores. The data within the FASTQ file is shown in Figure 1.3.

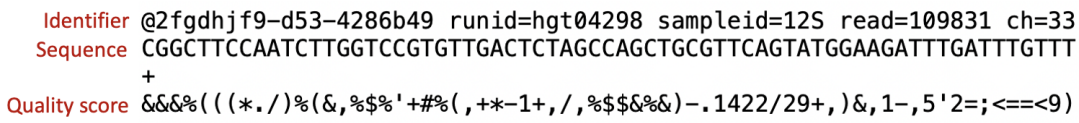


Figure 1.3: Raw Sequencing Data (FASTQ File)

The next step is to align these generated reads to a reference genome using an alignment algorithm such as Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009). The alignment process maps the reads to their corresponding genomic locations, taking into account potential variations, such as single nucleotide variants (SNVs) and small insertions or deletions (indels). After alignment the data is stored in a SAM/BAM (Sequencing/Binary Alignment Map) format. Once the reads are aligned, variant calling algorithms, such as the Genome Analysis Toolkit (GATK) (McKenna et al., 2010), are used to identify genetic variants within the exonic regions. Once the reads are aligned and variant calling algorithms are applied, the resulting data structure typically consists of a Variant Call Format (VCF) file. Table 1.1 shows a sample representation of the data structure after variant calling using GATK.

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chr1	10012	.	A	G	100	PASS	DP=50;AC=2;AN=2
chr1	10145	rs123456	C	T	500	PASS	DP=60;AC=1;AN=2;AF=0.5
chr2	22003	.	G	A	80	PASS	DP=40;AC=1;AN=1

Table 1.1: WEX: Variant Calling Output

In Table 1.1, each row represents a variant call at a specific genomic position (POS) on a

particular chromosome (CHROM). The ID column contains the unique identifier for the variant, while REF and ALT represent the reference and alternate alleles, respectively. QUAL represents the variant quality score, and FILTER indicates whether the variant passed quality filters. The INFO column provides additional information about the variant, such as read depth (DP), allele count (AC), allele number (AN), allele frequency (AF), and other annotations.

WEX data can also be used for Copy number alteration calling using the tools and algorithms that are discussed in Chapter 2. In Table 1.2, we show an outcome after copy number calling on WEX data. Each row represents a specific region (segment) on a chromosome (Chrom) with its corresponding start and end positions. In a diploid organism, most genes exist in two copies, one inherited from each parent. These copies can be identical (homozygous) or different (heterozygous). The major copy number refers to the more frequent or higher number of copies, while the minor copy number refers to the less frequent or lower number of copies. The Total_cn column indicates the estimated total copy number for that region, while Minor_cn and Major_cn represent the minor and major copy number estimates, respectively.

Chrom	Start	End	Total_cn	Minor_cn	Major_cn
chr1	10010	20123	4	1	3
chr1	20678	120345	3	1	2
chr1	120678	560345	5	2	3

Table 1.2: WEX: Copy Number Calling Output

In the clonality analysis, we incorporate data obtained from both Variant Calling algorithms and Copy Number Alteration (CNA) calling algorithms. Specifically, we integrate information from SNV mutations identified at specific genomic locations with the corresponding copy number estimates. When analyzing a particular genomic location that harbors a SNV mutation, we take into account the associated copy number estimate, which may deviate from the default value of

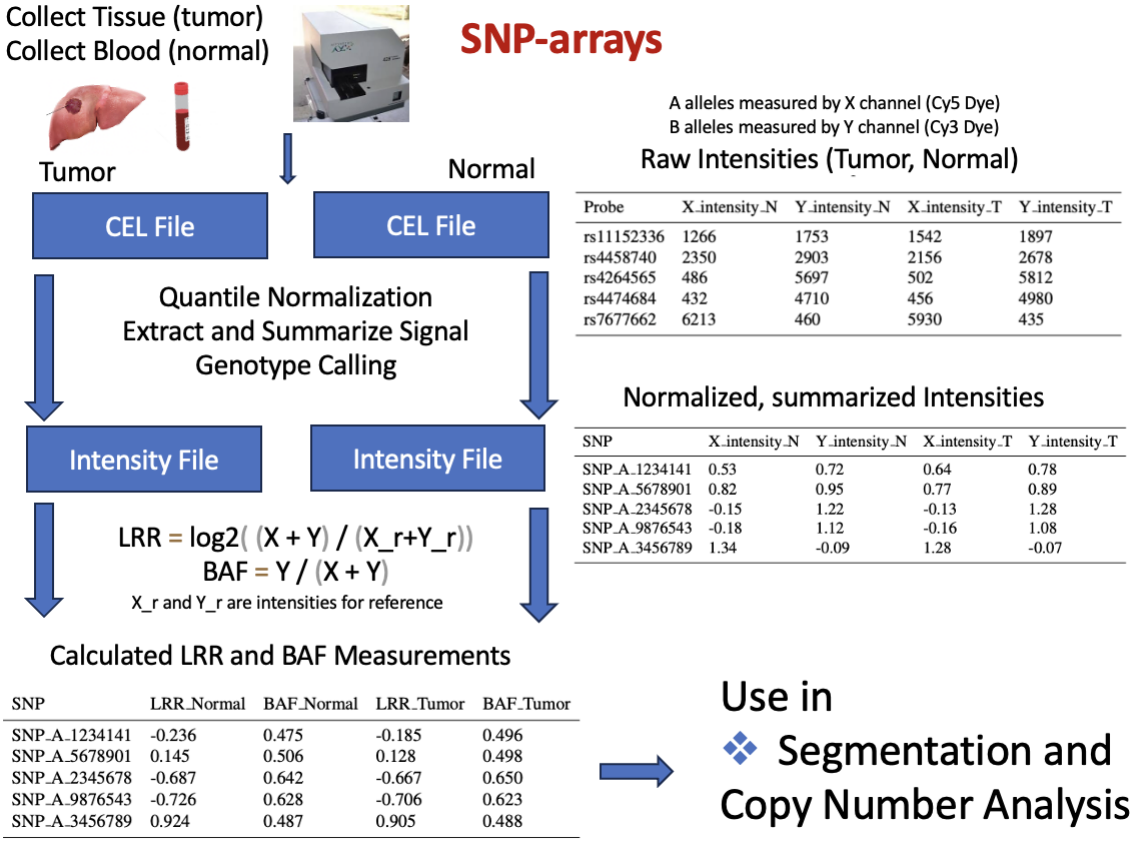
2 for diploid regions. By considering the copy number alteration information at that specific location, we aim to refine the accuracy of our clonality analysis results. This integration of copy number estimates with SNV mutations allows for a more comprehensive assessment of the clonal architecture and genetic aberrations within the tumor samples.

1.4 Single Nucleotide Polymorphism (SNP) Array Data

SNP (Single Nucleotide Polymorphism) arrays are another type of genomic data that are commonly used to investigate genetic variations in populations (LaFramboise, 2009). SNPs are single base pair variations in the DNA sequence that differ between individuals. In SNP-array experiments, probes are designed to specifically bind to and analyze specific genetic variants, such as SNPs. As shown in the workflow Figure 1.4, the raw output from array experiment will be the .CEL file which contains raw intensity measurements for each probe on the array X intensities corresponding to A alleles measured by X channel (Cy5 Dye, far-red fluorescent) and Y intensities corresponding to B alleles measured by Y channel (Cy3 Dye, orange fluorescent).

After obtaining the raw .CEL data in SNP array experiments, the next steps typically involve data preprocessing, quality control, normalization, and downstream analysis. The raw intensity measurements from the .CEL file are preprocessed to correct for systematic biases and artifacts. Quality control measures are applied to ensure the reliability and accuracy of the data. To account for technical variations between arrays, normalization methods are applied to bring the data onto a common scale. One of the key objectives in SNP array analysis is to determine the genotypes for each SNP. Genotype calling algorithms are used to assign genotypes (e.g., AA, AB, BB) based on the intensity measurements and reference data. These algorithms consider

factors such as allelic frequencies and clustering patterns to make accurate genotype assignments. After genotype calling, additional quality filters may be applied to remove low-confidence or ambiguous genotypes.



Probe	X_intensity_N	Y_intensity_N	X_intensity_T	Y_intensity_T
rs11152336	1266	1753	1542	1897
rs4458740	2350	2903	2156	2678
rs4264565	486	5697	502	5812
rs4474684	432	4710	456	4980
rs7677662	6213	460	5930	435

SNP	X_intensity_N	Y_intensity_N	X_intensity_T	Y_intensity_T
SNP_A_1234141	0.53	0.72	0.64	0.78
SNP_A_5678901	0.82	0.95	0.77	0.89
SNP_A_2345678	-0.15	1.22	-0.13	1.28
SNP_A_9876543	-0.18	1.12	-0.16	1.08
SNP_A_3456789	1.34	-0.09	1.28	-0.07

SNP	LRR_Normal	BAF_Normal	LRR_Tumor	BAF_Tumor
SNP_A_1234141	-0.236	0.475	-0.185	0.496
SNP_A_5678901	0.145	0.506	0.128	0.498
SNP_A_2345678	-0.687	0.642	-0.667	0.650
SNP_A_9876543	-0.726	0.628	-0.706	0.623
SNP_A_3456789	0.924	0.487	0.905	0.488

Figure 1.4: SNP Array Data Workflow

Finally, Log R Ratio (LRR), a measure of the relative DNA copy number at a specific SNP locus, and B allele frequency (BAF), the allelic composition of a SNP locus, are calculated. LRR is calculated by taking the log base 2 of the ratio of the observed intensity (the total intensity, $X + Y$) of the sample to the expected intensity (reference total intensity, $X_r + Y_r$). BAF is calculated by dividing the intensity of one of the alleles (usually referred to as the B allele) by the total intensity as shown in Figure 1.4.

In this thesis, the final SNP-array outcome data we utilize follows a format where rows

represent SNP locations and columns represent LRR and BAF measurements for both normal and tumor samples of each individual. Similar to the approach used in WEX copy number calling, we use these results with WEX variant calling data to enhance our clonality analysis in Chapter 3.

1.5 RNA Sequencing (RNA-seq)

RNA-seq is a powerful genomic technology used to measure gene expression levels of RNA molecules ((Wang et al., 2009)). In this thesis, we primarily focus on bulk RNA-seq data and do not utilize single-cell RNA-seq. In bulk RNA-seq, RNA is extracted from a mixed population of cells, and the resulting data represents the average expression profile of all cells in the sample. This data is organized in a count matrix format, where rows correspond to genes and columns represent samples, such as biological replicates or technical replicates. Each entry in the matrix represents the number of reads mapped to a specific gene in a particular sample. Unlike bulk RNA-seq, single-cell RNA-seq captures gene expression profiles of individual cells within a sample. Each cell is isolated, and its RNA is sequenced separately. The resulting data also follows a count matrix format, but the rows now represent genes, and the columns represent individual cells. The matrix contains the counts of RNA molecules originating from each gene in each individual cell.

The process of RNA-seq involves several steps, starting with the isolation of RNA from a biological sample. The RNA is then reverse transcribed into complementary DNA (cDNA) . This cDNA is fragmented, and sequencing adapters are added to create a library of fragments ready for sequencing (Kukurba and Montgomery, 2015). The generated short sequencing reads capture

information about the expression levels and sequence composition of RNA transcripts. These reads are aligned to a reference transcriptome using alignment algorithms such as HISAT, STAR, or TopHat (Kim et al., 2015).

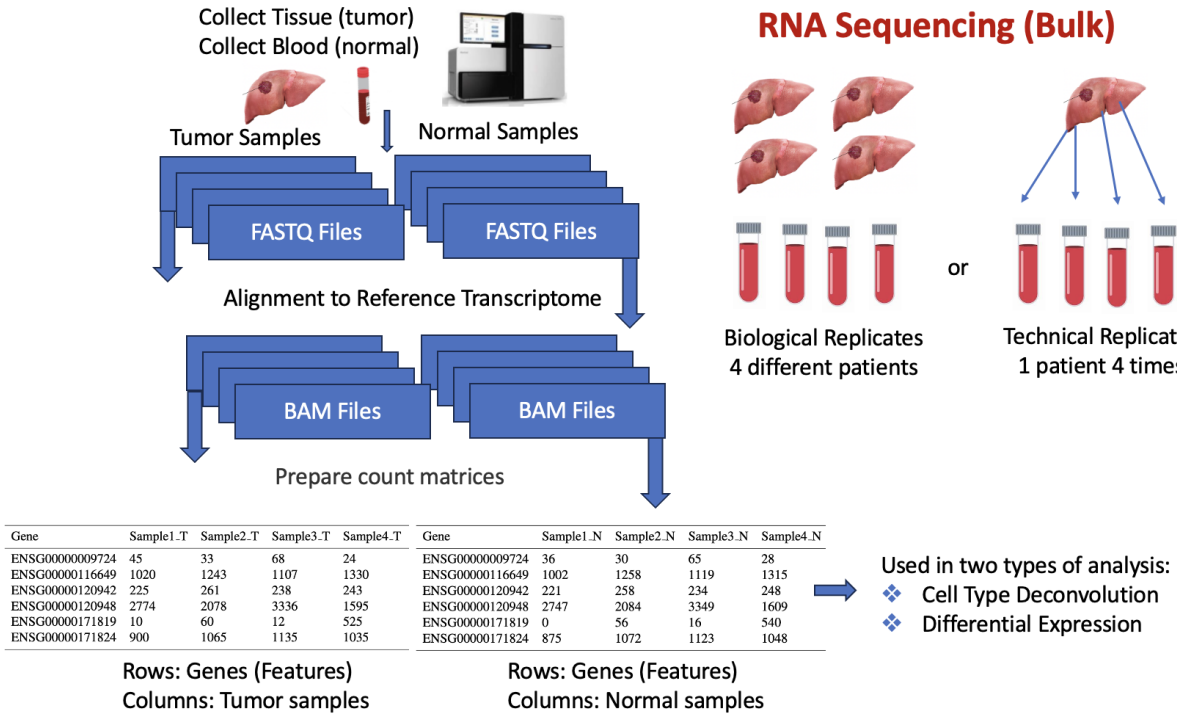


Figure 1.5: RNA-sequencing Workflow

Quantification of transcript abundance involves assigning reads to specific genes or transcripts. This step includes counting the number of reads that align to each gene or transcript, resulting in a measure of expression level. Popular quantification methods include featureCounts and HTSeq (Anders et al., 2015; Liao et al., 2014). The quantified data is then organized in a count matrix, as shown in Figure 1.5, where each gene is associated with its respective expression counts across different samples (either tumor samples from a cohort or normal samples from a cohort).

Differential expression analysis is commonly performed to compare gene expression levels between different conditions or groups (e.g., normal versus tumor). Statistical methods such as edgeR, DESeq2, or limma are used to identify genes that show significant differences in

expression between groups (Li, 2019). The results of the differential expression analysis are typically presented in a table format, as shown in Table 1.3, including information such as the log2 fold change, which refers to the ratio of expression levels between two conditions (e.g. tumor vs normal), adjusted p-value, and significance of differential expression.

Gene	Log2 Fold Change	Adjusted p-value	Significant	Sample 1 Mean	Sample 2 Mean
Gene 1	2.5	0.001	Yes	1000	500
Gene 2	-1.8	0.025	Yes	800	900
Gene 3	0.2	0.852	No	1500	700
Gene 4	-3.1	0.001	Yes	400	300

Table 1.3: RNA-seq Data: Differential Expression Analysis Results

In Table 1.3, the "Log2 Fold Change" column provides fold change, ratio of the expression level of tumor to the expression level of normal. The "Adjusted p-value" column represents the statistical significance of the differential expression, adjusted for multiple testing. Significant indicates whether the gene is significantly differentially expressed (e.g., based on a chosen threshold or adjusted p-value cutoff). In this table, one count for each gene type is demonstrated for illustration purpose (Gene1, Gene2, etc.). It is important to note that in general for RNA-seq data, it is possible to have multiple variants or isoforms of a gene within a sample tissue. Additionally, alternative splicing, a process where different combinations of exons are selected during mRNA (messenger RNA) processing, can further contribute to the diversity of gene expression patterns.

In this research, we utilize the count matrix for tumor samples shown in Figure 1.5 for each cohort. These gene counts are used in downstream analysis of cell type abundance described in Section 3.4.3.1.

1.6 Copy Number Alterations (CNAs) and Single Nucleotide Variants (SNVs)

CNAs and SNVs are common genomic alterations in cancer cells that are indications of tumor heterogeneity. CNAs refer to changes in the number of copies of specific genomic regions, while SNVs are point mutations in individual nucleotides.

The detection and characterization of CNAs and SNVs rely on advanced genomic technologies. Array Comparative Genomic Hybridization (aCGH) and Single Nucleotide Polymorphism (SNP) arrays enable genome-wide profiling of copy number alterations and identification of genetic variations, respectively. Whole Exome Sequencing (WEX) provides a comprehensive assessment of coding regions, allowing for the detection of SNVs and small insertions or deletions. These techniques, combined with sophisticated bioinformatics analyses, provide researchers with a wealth of information to explore tumor heterogeneity, clonal evolution, and potential driver events. Integrating CNAs and SNVs data allows for a more comprehensive understanding of the complex genomic landscape of cancer.

1.7 Liver Cancer Cohorts

This study aims to analyze three independent liver cancer cohorts: TIGER-LC (Chaisaingmongkol et al., 2017), NCI-MONGOLIA (Candia et al., 2020), and TCGA-LIHC (The Cancer Genome Atlas Research Network, 2017). Each cohort has its own unique characteristics, which are detailed in the accompanying table. Both WEX and RNA-seq data for most of the patients in these cohorts are available. Currently, liver cancer is categorized into two primary forms: hepatocellular carcinoma (HCC) and cholangiocarcinoma (CCA) . However, extensive

research has shown that within these forms, there are diverse molecular subtypes of tumor cells, characterized by mutation and specific genetic drivers and clinical outcomes. It is worth noting that the TIGER-LC cohort is particularly intriguing as it primarily consists of CCA tumors, despite HCC being the predominant type globally, accounting for 90% of cases. Initially, the analysis will focus on patients with HCC tumors, as the TCGA-LIHC and NCI-MONGOLIA cohorts exclusively comprise HCC cases. Subsequently, we will delve deeper into the TIGER-LC cohort as a whole.

During the analysis of the survival data, an examination is carried out to explore all available information concerning the three cohorts, including details on the last contact reasons for the subjects. However, it is important to acknowledge that there is a lack of explicit information regarding whether the right-censoring of patients represented end-of-study events or instances of dropout.

Variable	TCGA-LIHC	TIGER-LC	NCI-MONGOLIA
	(WEX:375, RNA:369)	(WEX:78, RNA:51)	(WEX:71, RNA:69)
Age, $\leq$60 y.o.	194	20	33
Age, $>$60 y.o.	180	56	37
NA	1	2	1
Gender, male	254	59	36
Gender, female	121	17	34
NA	0	0	1
Stage, early	261	23	19
Stage, late	90	10	35
NA	24	45	17
Survival status, alive	244	37	49
Survival status, dead	131	39	22

Table 1.4: Clinical information of liver cancer cohorts
WEX: Whole-Exome sequencing. RNA: RNA sequencing.

TCGA LIHC: The Cancer Genome Atlas (TCGA) hepatocellular carcinoma (TCGA LIHC)

annotated mutation and RNA sequencing data files for 378 cases were extracted from the GDC (Genomic Data Commons Data Portal) (Grossman et al., 2016). There were 2 tumor samples for 1 patient and 3 tumor samples for another patient, so we only included 1 tumor sample from each patient to do further analysis. We opted to include only the initial tumor samples from these patients in our study.

NCI MONGOLIA: HCC patients were diagnosed via standardized pathology reviews based on the WHO Classification of Tumors (also known as the WHO Blue Books) and via clinical assessments based on CT scans and ultrasound diagnosis. Tumoral and adjacent non tumoral liver tissue samples were collected and frozen at 20 C° after surgical resection at the National Cancer Center in Ulaanbaatar, Mongolia. The study was approved by the Ethics Committee at the National Cancer Center in Ulaanbaatar, Mongolia, and written informed consent was obtained from all participants.

TIGER LC: A set of surgical specimens from 78 HCC patients were used. Patients were diagnosed using combinations of imaging studies, tumor size, level of alpha-fetoprotein (AFP) and histological investigations. Informed consent was obtained from all patients included in this study and approved by the Institutional Review Boards of the respective institutions (NCI protocol number 13CN089; CRI protocol number 18/2555; Chulabhorn Hospital protocol number 11/2553; Thai NCI protocol number EC163/2010; Chiang Mai University protocol number TIGER LC; Khon Kaen University protocol number HE541099).

1.8 Tumor evolution trajectories

Tumor evolution refers to the process by which cancer cells accumulate genetic changes over time, leading to the development of distinct subclones within the tumor (Nowell, 1976). Understanding the evolution trajectory of a tumor is important for developing effective personalized treatment strategies for cancer patients. In cancer research, several different evolution trajectories have been observed, including linear, branching, and neutral evolution (Davis et al., 2017).

Linear evolution occurs when a single clone of cancer cells acquires additional mutations over time, leading to more aggressive tumors. This trajectory has been observed in liver cancer and is associated with poor prognosis (Zhu et al., 2021).

Branching evolution occurs when multiple subclones of cancer cells emerge from the original tumor, each with distinct genetic profiles. This can lead to the development of multiple tumors or metastases in different parts of the body. Branching evolution has also been observed in liver cancer and is associated with a more aggressive disease course and reduced survival rates (Castelli et al., 2017).

Neutral evolution in tumors refers to a scenario where various clones of cancer cells coexist without any single clone having a significant advantage over others. These clones accumulate mutations, but these genetic changes do not provide a selective benefit in terms of cell survival or growth. As a result, the tumor exhibits similar sizes and lacks a dominant clone. Additionally, tumors driven by neutral evolution may display limited response to targeted therapies since there is no dominant clone with specific vulnerabilities to exploit. In liver cancer, neutral tumor evolution has been proposed as a possible explanation for the heterogeneity observed in this disease, highlighting the complex nature of tumor development and progression (Craig et al.,

2020).

Personalized medicine aims to develop treatment strategies that are tailored to the specific genetic profile of a patient's tumor. By understanding the evolution trajectory of a tumor, oncologists hope to develop personalized treatment strategies that target the specific genetic mutations driving the growth and spread of the cancer.

1.9 Clustering algorithms: application to genomic data

Clustering algorithms are widely used in the analysis of genomic data to identify underlying patterns and group similar entities together. These algorithms partition the data into distinct clusters based on the similarity or dissimilarity between the entities.

1.9.1 Similarity and dissimilarity measures: where do they come from?

Before delving into the various clustering methods, it is important to understand the concept of similarity and dissimilarity measures. Similarity is a quantitative measure that indicates the degree of resemblance or likeness between two data records or objects. It is used to assess the similarity between samples in clustering algorithms. A higher similarity value suggests that the samples are more similar or alike. Dissimilarity is the opposite of similarity and measures the degree of difference or dissimilarity between two data records or objects. It quantifies the dissimilarity between samples in clustering algorithms. A lower dissimilarity value indicates that the samples are more dissimilar or different. Distance measures are mathematical functions used to quantify the dissimilarity between pairs of samples in clustering algorithms. They are often used interchangeably with dissimilarity measures, although it is worth noting that not all

dissimilarity measures are strictly distance functions.

Distance measures can take various forms, depending on the nature of the data and the specific requirements of the clustering algorithm. Commonly used distance measures include:

Euclidean distance: This is one of the most widely used distance measures, especially for numerical data. It calculates the straight-line distance between two data points in the feature space.

Manhattan distance: Also known as city block distance or L_1 norm, this distance measure calculates the sum of absolute differences between the coordinates of two data points along each dimension.

Cosine similarity: While not a distance measure per se, cosine similarity is often used as a dissimilarity measure for high-dimensional data. It calculates the cosine of the angle between two data vectors and provides a measure of their similarity.

Pearson correlation coefficient: Primarily used for measuring the linear relationship between two numerical variables, the Pearson correlation coefficient can also be employed as a dissimilarity measure. It measures the strength and direction of the linear association between two variables.

Hamming distance: This distance measure is commonly used for binary or categorical data. It calculates the number of positions at which two data records differ.

Jaccard distance: Often used for set-like data, the Jaccard distance quantifies the dissimilarity between two sets by considering the size of their intersection and union (Jaccard, 1912).

Dissimilarity measures can be derived from various sources depending on the nature of the data and the specific context of the clustering problem. Here are a few examples:

Domain-specific knowledge: In some cases, domain experts have prior knowledge about the data and can design dissimilarity measures based on specific characteristics relevant to the problem at

hand (Bilal, 2021). For example, in genomic clustering, dissimilarity measures can be based on the differences in gene expression levels, genetic variants, or other genomic features.

Feature engineering: Dissimilarity measures can be derived by transforming or combining features to capture the dissimilarity between samples. This process is known as feature engineering, where domain knowledge and statistical techniques are used to create informative and relevant dissimilarity measures. For instance, in text clustering, dissimilarity measures can be based on word frequency, document similarity, or other text features.

Statistical measures: Statistical techniques can be used to compute dissimilarity measures based on the statistical properties of the data. For example, in time-series clustering, dissimilarity measures can be derived from statistical measures such as correlation coefficients, autocorrelation, or spectral analysis.

Information theory: Dissimilarity measures can be based on concepts from information theory, such as entropy or mutual information. These measures capture the information content or the degree of uncertainty between pairs of samples. Entropy measures are commonly used in the context of clonality analysis, specifically for estimating the diversity and heterogeneity of clonal populations.

Machine learning techniques: Machine learning algorithms, such as neural networks, decision trees or random forest, can be trained to learn dissimilarity measures directly from the data (Xing et al., 2002). These approaches, often referred to as metric learning or distance metric learning, aim to optimize a distance function that captures the dissimilarity between samples based on the clustering objectives. In this thesis, we use a method of defining similarity using random forests.

It is important to choose an appropriate distance or dissimilarity measure that aligns with the nature of the data and the objectives of the clustering task. Different distance measures may

yield different clustering results, so selecting the most suitable measure is crucial for obtaining meaningful clusters.

1.9.2 Common clustering methods for genomic data analysis

Genomic clustering aims to group individuals based on their genomic characteristics, which includes their DNA sequence, gene expression levels, genetic variants, and other genomic features. Clonal structure of tumor represented as a phylogenetic tree is one of the inferred outcomes obtained using these genomic features. In this work, we use clustering algorithms to obtain biologically meaningful and distinct phylogenetic tree groups. We next give brief definitions of types of clustering methods that are used in genomic clustering.

Hierarchical clustering: This class of methods builds a tree-like structure, known as a dendrogram, by iteratively merging or splitting clusters based on the similarity or dissimilarity between observations. AGNES (Agglomerative Nesting) (Kaufman and Rousseeuw, 1990) is an agglomerative hierarchical clustering algorithm. It starts by assigning each sample to its own cluster and then iteratively merges the closest clusters based on a distance measure until a stopping criterion is met. The choice of distance measure, such as Euclidean distance or Pearson correlation, plays a crucial role in AGNES clustering. DIANA (Divisive Analysis) (Kaufman and Rousseeuw, 1990) is a divisive hierarchical clustering algorithm, which takes the opposite approach of AGNES. It starts with all samples in a single cluster and recursively splits clusters into smaller groups until a stopping criterion is satisfied. DIANA uses various techniques, such as the divisive coefficient, to determine the optimal cluster divisions.

In the AGNES and DIANA hierarchical clustering algorithms, the distances between clusters

of multiple units are defined using linkage methods. These methods determine how the distances between pairs of units within respective clusters are combined to compute the distance between clusters. The most commonly used linkage methods include maximum, minimum, and average linkage. In maximum linkage, also known as complete linkage, the distance between two clusters is defined as the maximum distance between any pair of units from each cluster. It measures the dissimilarity between clusters based on the most distant pair of units. This linkage method tends to preserve compact and well-separated clusters. In minimum linkage, also known as single linkage, the distance between two clusters is defined as the minimum distance between any pair of units from each cluster. It measures the dissimilarity between clusters based on the closest pair of units. This linkage method tends to form long, chain-like clusters and can be sensitive to noise and outliers. In average linkage, the distance between two clusters is defined as the average distance between all pairs of units, where one unit is from each cluster. It measures the dissimilarity between clusters based on the average similarity between their units. This linkage method provides a balanced approach, considering all pairs of units. Other linkage methods, such as centroid linkage and Ward's linkage, are also commonly used in hierarchical clustering algorithms. Centroid linkage calculates the distance between clusters based on the centroids (mean vectors) of the units within each cluster. Ward's linkage minimizes the increase in within-cluster variance when merging clusters.

The choice of linkage method may impact the resulting clustering structure. Each method has its own characteristics and can lead to different cluster shapes and sizes. Researchers should carefully consider the nature of the data and the objectives of the analysis when selecting an appropriate linkage method.

Partitioning based clustering: K-means clustering (Hartigan and Wong, 1979) is a popular

and widely used partitioning-based clustering algorithm. It aims to partition samples into k clusters, where k is a predefined number. The algorithm begins by randomly initializing k cluster centroids. It then iteratively assigns each sample to the nearest centroid based on a distance measure, commonly the Euclidean distance. After all samples have been assigned, the centroids are updated by calculating the mean of the samples assigned to each cluster. This assignment and centroid updating process is repeated until convergence, where the assignments and centroids no longer change significantly. The final result is a set of k clusters, each represented by its centroid. K-means is efficient and scalable, making it suitable for large datasets. However, it is sensitive to the initial centroid placements and may converge to local optima. Multiple runs with different initializations are often performed to improve the clustering solution. PAM (Partitioning Around Medoids) (Kaufman and Rousseeuw, 1990) is another partitioning-based clustering algorithm that extends the concept of k-means by using medoids as representatives of the clusters instead of centroids. Medoids are actual samples from the dataset and are chosen to minimize the average dissimilarity to other samples in the same cluster. The algorithm starts by selecting k initial medoids randomly or using a more sophisticated initialization technique. It then iteratively evaluates the cost of swapping a medoid with a non-medoid sample and performs the swap if it reduces the total dissimilarity within the cluster. PAM repeats this process until no further improvement can be made. PAM is more robust to noise and outliers compared to k-means because it uses actual samples as medoids. However, it can be computationally expensive, especially for large datasets, as it requires evaluating pairwise dissimilarities between samples.

The success of partitioning-based clustering methods depends on selecting the appropriate number of clusters (k) and the choice of distance measure. Determining the optimal value of k is often a challenge, and there are various approaches, such as the elbow method or silhouette

analysis (Rousseeuw, 1987; Tibshirani et al., 2001), to help determine the optimal number of clusters. The choice of distance measure should align with the characteristics of the genomic data and the research objectives.

1.9.3 Bootstrapping method for assessing clustering stability

In this research, we employ a bootstrapping method to assess the stability of the clustering results obtained from the clonality analysis. Bootstrapping is a resampling technique that allows us to estimate the variability and uncertainty associated with a statistical analysis by repeatedly sampling from the original dataset with replacement. This method is particularly useful in evaluating the stability of clustering algorithms and their outcomes.

The primary objective of using bootstrapping in the context of clustering analysis is to assess the consistency of the identified clusters when the dataset is perturbed. By generating multiple bootstrap samples, each consisting of randomly selected observations from the original dataset, we can investigate the stability of the clusters across different iterations of the resampling process.

The following steps outline the general procedure for applying the bootstrapping method to study the stability of clustering results:

Bootstrap sampling: Randomly select a subset of observations from the original dataset, with replacement, to create a bootstrap sample. The size of the bootstrap sample is typically the same as the original dataset, but with some observations being replicated and others omitted.

Clustering analysis: Apply the clustering algorithm, such as the random forest clustering algorithm used in this research, to the bootstrap sample. Obtain the clustering results, including the assignment

of observations to clusters.

Repeat: Repeat steps 1 and 2 a large number of times, generating multiple bootstrap samples and performing clustering analysis on each sample.

Cluster agreement calculation: Calculate the cluster agreement or similarity between different bootstrap samples. This can be done using various metrics, such as the Jaccard index or affinity score (used in this work), which measure the similarity between two sets of clusters.

Assess cluster stability: Analyze the cluster agreement across multiple bootstrap samples to assess the stability of the clustering results. Higher agreement or similarity values indicate more stable clusters, while lower values suggest instability or variability in the cluster assignments.

Interpretation and validation: Interpret the cluster stability results and validate the robustness of the identified clusters. Compare the stability results with the original clustering analysis to evaluate the consistency and reliability of the clusters.

1.10 Comparing survival distributions of defined groups: Log-rank Tests

Cancer research is a dynamic and vital field aimed at understanding the complexities of cancer progression, treatment response, and patient outcomes. Survival analysis serves as a fundamental statistical methodology in this domain, enabling the investigation of time-to-event data such as overall survival, disease-free survival, and recurrence-free survival times (possibly right-censored) for individual patients. In survival analysis, assumptions regarding the censoring mechanism are necessary to bridge the gap between the observed data and the underlying phenomenon of interest, particularly in the presence of right censoring. Noninformative censoring is an assumption in survival analysis that assumes the conditional independence between

failure time and censoring time, given the treatment group assignment and covariates. In this thesis, we adopt this assumption, among others discussed in the literature (Andersen and Gill, 1982; Kaplan and Meier, 1958; Overgaard and Hansen, 2019). The detailed discussion and elaboration of these assumptions can be found in Chapter 4. Within the realm of survival analysis, the log-rank test is one of the most popular techniques, allowing for the comparison of survival curves between different groups or treatment arms. The test determines whether there are significant differences in the survival curves of different groups based on the observed event times.

The log-rank test (Bland and Altman, 2004; Peto and Peto, 1972) compares the accumulated observed number of events in each group to the conditional-expected number of events. These counts are calculated over finely spaced disjoint time intervals, taking into account the survival probabilities or cumulative hazard functions estimated from the data. The log-rank statistic is used to test the null hypothesis of no difference in survival between the groups, providing a statistical evaluation of the observed versus expected event rates.

K-sample log-rank test is a variant of the log-rank test that specifically compares the survival distributions among multiple groups simultaneously. It is used when there are more than two groups being compared. The test examines whether there are any significant differences in survival patterns among the groups. The K-sample log-rank test is an extension of the log-rank test that accommodates multiple treatment or exposure groups.

The stratified log-rank test (Peto et al., 1976) is an extension of the basic log-rank test that accounts for stratification of a study population into homogeneous but distinct groups. The idea of stratification is the anticipation that any non-null survival differences will consistently occur in the same direction across all stratum-groups. In other words, if there are true differences in survival between groups, we expect these differences to be present in the same direction but quantitatively

different degrees across all subgroups defined by the stratification variables. By stratifying, we enhance our ability to detect these meaningful differences in survival. By stratifying the data based on these variables, the stratified log-rank test compares the survival distributions between groups within each stratum separately. The stratified log-rank test calculates a separate log-rank statistic for each stratum and combines them into an overall test statistic.

The covariate-adjusted log-rank test (Kong and Slud, 1997; Ye et al., 2023) is another extension of the log-rank test that incorporates additional covariates into the analysis. Covariates are variables that have the potential to influence survival outcomes, irrespective of the grouping variable under investigation. However, in the context of survival analysis, the concern arises when there is covariate imbalance. This imbalance refers to situations where covariates affect survival but in a manner that does not exhibit differences between the treatment groups under the null hypothesis of no survival difference. Despite the absence of apparent differences in covariate effects between treatment groups under the null hypothesis, these covariates can still play a crucial role in detecting survival differences when the treatment groups exhibit divergent survival patterns. In such cases, the inclusion of covariates becomes invaluable as they provide additional information that can uncover and elucidate the survival disparities between the treatment groups. By adjusting for these covariates, the covariate-adjusted log-rank test allows for a more precise evaluation of the effect of the grouping variable on survival, taking into account the influence of the covariates.

Overall, these modifications of the log-rank test provide more powerful assessments of the differences in survival between groups by considering stratification variables and covariates that may influence the outcomes. They offer valuable tools for validating and analyzing survival outcomes in various research settings.

1.11 Validating genomic clustering through survival analysis

The idea is to use survival information as a form of external validation to assess the validity and clinical significance of the constructed phylogenetic clusters in liver cancer. Hypothesis testing techniques, specifically utilizing the log-rank test statistic, are employed to evaluate and validate these clusters.

By applying the log-rank test, we compare the survival distributions among the different genomic clusters. This statistical test helps us determine if there are significant differences in survival outcomes between the clusters. If the log-rank test yields statistically significant results, it provides evidence that the membership in genomic clusters has a significant impact on patient survival. This suggests that the identified clusters are associated with distinct prognosis and can potentially serve as important prognostic indicators.

By using survival information as an external validation source, we can assess the clinical relevance and validity of the genomic clusters derived from phylogenetic analysis. Through hypothesis testing with the log-rank test and its modifications, we obtain rigorous statistical evidence to support the hypotheses associated with the clusters. This validation process enhances our understanding of the clinical implications of the identified clusters and helps guide clinical decision-making in liver cancer treatment and management.

1.12 Summary of contributions

This thesis makes several contributions to the field of genomic analysis and cancer research, with some methodological advances in mathematical statistics. The main contributions are the

following.

1. Copy number tools comparison:

   This contribution focuses on the comparison of copy number analysis tools. An overview of various copy number analysis tools is presented. The strengths, limitations, and underlying principles of popular methods are discussed. The selected copy number analysis tools are applied to the datasets in this research and comparison results are provided. Through these comparisons, researchers can gain understanding into the strengths and limitations of these tools, helping them choose the most appropriate algorithm for their specific copy number alteration analysis tasks.

2. Clonality analysis with clustering algorithm:

   This contribution introduces an implementation of a clonality analysis approach using a unique clustering algorithm, aligning with the steps mentioned in the pipeline Figure 1.1. A framework is proposed for clonality analysis using subclone proportion and mutation probability vectors. This framework allows for the identification of biologically meaningful clusters in tumor evolution. The utilization of a clustering algorithm on nonlinear tumor trees successfully categorizes the trees into distinct groups. The stability of the identified clusters is assessed using the bootstrap method, as outlined in the pipeline. Multiple bootstrap samples are generated, and the clustering algorithm is applied to evaluate the consistency and stability of the resulting clusters. Additionally, data displays such as principal component analysis (PCA) and scatter plots are employed to visualize the distribution and overlap of the clusters in a reduced-dimensional space. Findings into tumor evolution dynamics and the heterogeneity of tumor populations are provided based on the identified

clusters.

3. Survival testing as a method of cluster validation:

   In this contribution, survival testing is employed to validate the identified clusters from the clonality analysis. The log-rank test and its modifications, widely employed statistical techniques in survival analysis, are applied to genomic data to assess the impact of cluster membership on patient survival. The methodological definition of the log-rank test and its modifications is presented. The log-rank test, incorporating covariates and multi-sample stratification, is used to evaluate the survival differences between the clusters. Furthermore, a methodological advance is made by extending the covariate-adjusted log-rank testing to multiple treatment groups with a K-sample covariate-adjusted log-rank statistic. This extension enhances the cluster validation process and contributes to the methodological advancement of survival analysis, applicable beyond the specific domain of cancer research. The validation using survival testing strengthens the assessment of the identified clusters in the context of tumor evolution.

Overall, this thesis not only contributes to cancer genomics and genomic analysis but also provides methodological advancements in applied clustering methodology and survival analysis, which have broader implications and applicability beyond the field of cancer research.

Chapter 2: Copy number alteration (CNA) detection tools

2.1 Overview

CNA (Copy Number Alteration) is a critical genomic feature observed in cancer cells, which can contribute to their progression and aggressiveness. It plays a crucial role in allowing accurate counting of the number of mutant reads in processed WEX (Whole Exome Sequencing) data. The detection of CNA is facilitated by various genomic data sets, including whole exome sequencing (WEX) data and SNP (Single Nucleotide Polymorphism) array data. In this thesis, we focus on the use of WEX data for the identification and analysis of CNA in tumor cells although we also utilize SNP array data of one cohort for comparison purposes. A SNP array utilizes SNP marker probes designed to specifically target particular genomic locations (Lin et al., 2013). Each SNP locus is assigned two distinct probes to target two alleles. The measurement of signal intensities involves assessing the combined hybridization intensities of these two probes. The Log R ratio (LRR), which serves as a normalized measure of signal intensity, is computed by taking the base-2 logarithm of the ratio between the observed signal and the expected signal for two copies of the genome at each SNP marker (de Araújo Lima and Wang, 2017). The B Allele Frequency (BAF) is an inference of the relative ratio of fluorescent signals between two probes or alleles at each SNP marker. It provides information about the allelic composition and can be used to assess the presence of genetic variants or copy number alterations.

Tool	Explanation	Language	Data Type	Method
ASCAT (2010)	ASCAT (Allele-Specific Copy number Analysis of Tumors) is a tool that estimates allele-specific copy number profiles from SNP array data using a Bayesian hierarchical model.	MATLAB, R	SNP array	PCF
Sequenza (2015)	Sequenza is a tool designed for inferring tumor purity, ploidy, and copy number profiles from tumor-normal paired whole-exome sequencing data. It utilizes a combination of statistical algorithms based on binomial mixture models.	Python, R	WEX	PCF
FACETS (2016)	FACETS (Fraction and Allele-Specific Copy Number Estimates from Tumor Sequencing) is a tool that estimates tumor purity, ploidy, and allele-specific copy number profiles from tumor-only DNA sequencing data.	R	WEX	CBS
hsegHMM (2018)	hsegHMM is an R package that performs segmentation and profiling of allele-specific copy number alterations using Hidden Markov Models. It utilizes the Viterbi algorithm and a maximum likelihood approach to identify genomic regions with altered copy numbers.	R	WEX	HMM

Table 2.1: Copy Number Alteration Analysis Tools

Whole exome sequencing (WEX) data provides the number of reads that map to targeted genomic regions (exons) that overlap with a sliding window used in sequencing (Zhao et al., 2020). After some filtering, trimming and normalization steps, segmentation is applied to aligned sequence data. The segmentation step utilizes statistical models to detect CNA regions. These models operate on the WEX data in BAM format, which includes read count information at known SNP locations. These read counts are used to estimate the signal intensity value Log R ratio (LRR) and B Allele Frequency (BAF) that are normally obtained from SNP array data. The algorithms are individually applied to multiple distinct genomic regions, each corresponding to either a chromosome or a chromosome arm (for long chromosomes). There are three approaches

commonly used in the segmentation step: Piecewise Constant Fitting (PCF) algorithms, Circular Binary Segmentation (CBS) algorithms, and Hidden Markov Model (HMM) based algorithms.

The aim of this chapter is to review and compare existing CNA detection tools (the ones we applied to the liver cancer data), given the current lack of a gold standard for performance evaluation. Through this review and comparison, this chapter aims to provide some background on the usage of existing CNA detection tools.

2.2 Piecewise Constant Fitting (PCF) algorithms

Piecewise Constant Fitting (PCF) algorithms (Nilsen et al., 2012) are computational methods that aim to identify and characterize CNAs by segmenting the genome into regions with distinct copy number states.

In a normal situation, the copy number profile is expected to be relatively constant for most genes across the genome. However, in cancer cells, genetic alterations can occur, leading to changes in the copy number profile. The main idea behind PCF algorithms is to approximate the copy number profile of a tumor sample as a piecewise constant function. This means that the genome is divided into consecutive segments, and within each segment, the copy number is constant. By estimating the copy number values and segment boundaries, PCF algorithms provide a comprehensive representation of CNAs in the tumor genome. PCF algorithms typically follow a series of steps to identify CNAs:

1. Preprocessing: The input data, such as DNA sequencing or microarray data, is preprocessed to remove noise and correct for technical biases. This may involve background correction, normalization, and quality control steps.

2. Segmentation: The genome is divided into segments, and the copy number within each segment is assumed to be constant. Segmentation methods aim to identify the boundaries between these segments based on the copy number profiles. Various statistical approaches, such as change-point detection algorithms, are employed to detect significant changes in copy number.

3. Model fitting: Once the segments are identified, PCF algorithms fit a piecewise constant model to the copy number data within each segment. This involves estimating the copy number value for each segment based on statistical methods such as least squares regression or maximum likelihood estimation.

4. Post-processing: After fitting the model, post-processing steps may be applied to refine the segmentation and improve the accuracy of the copy number estimates. These steps can include noise reduction, outlier detection, and merging or splitting segments based on certain criteria.

5. Visualization and interpretation: The final output of PCF algorithms is a segmented copy number profile, which can be visualized using plots or heatmaps. Researchers can examine the identified CNAs and their boundaries to gain insights into the genomic alterations present in the tumor sample. Interpretation of the results often involves comparing the identified CNAs to known cancer-related genes or functional regions of the genome. Next, we discuss software tool, ASCAT (Van Loo et al., 2010) and Seqeunza (Favero et al., 2015) that utilize PCF methods for segmentation step of copy number estimation.

2.2.1 ASCAT Algorithm

The ASCAT (allele-specific copy number analysis of tumors) (Van Loo et al., 2010) model is a method used for detecting copy number alterations (CNAs) in tumor samples from SNP array

data. The model is based on the assumption that the copy number at a given genomic location can be expressed as a function of the allele-specific copy numbers n_A and n_B, where n_A denotes the number of copies of the A or wild type allele and n_B denotes the number of copies of the B allele. The SNP array data provides LRR (Log R Ratio) and BAF (B allele frequency) values, which are used to estimate the values of n_A and n_B. Sample data input of ASCAT is given in Table 2.2.

Name	Chr	Position	LRR.tumor	BAF.tumor	LRR.normal	BAF.normal
SNP_A-2131660	1	1145994	0.30	0.48	0.32	0.48
SNP_A-1967418	1	2224111	-0.52	0.87	-0.46	0.93
SNP_A-1969580	1	2319424	-0.59	1.00	-0.48	0.98
SNP_A-4263484	1	2543484	0.13	0.00	0.15	0.00
SNP_A-1978185	1	2926730	0.46	0.47	0.26	0.46
SNP_A-4264431	1	2941694	-0.19	0.54	-0.27	0.37

Table 2.2: Data Input for ASCAT Software

Here, LRR measures the the total intensity signals for both alleles, and the BAF is the relative proportion of one of the alleles with respect to the total intensity signal at each SNP locus. Because they provide complementary information, both LRR and BAF signals are required for a complete characterization of copy number changes and allelic ratio. We note that these input data are already an inference product, specifically, for each SNP marker with two alleles, the raw intensities for A and B alleles are subject to normalization and generate normalized intensity X and Y. Then, LRR and BAF values are calculated using the total intensity $R = X + Y$ and relative intensity $\theta = \frac{\arctan(Y/X)}{\pi/2}$. $LRR = \log 2(\frac{R_{observed}}{R_{expected}})$ where $R_{expected}$ is calculated as described in (Peiffer et al., 2006). BAF values are also calculated as in (Wang et al., 2007).

2.2.1.1 Background of data and model

The Log R Ratio (LRR) and B Allele Frequency (BAF) at a specific genomic location (SNP) can be represented as functions of the allele-specific copy numbers n_A and n_B. In the case of diploid and homogeneous samples, meaning the sample purity is 1, and when measurement noise is disregarded, the LRR and BAF values (referred to as r and b, respectively) can be accurately approximated by the following equations for i_{th} genomic location:

$$r_i = \gamma \log_2(\frac{n_{A,i} + n_{B,i}}{2}), \tag{2.1}$$

$$b_i = \frac{n_{B,i}}{n_{A,i} + n_{B,i}}. \tag{2.2}$$

When tumor ploidy is different than 2, this will cause a shift in LRR although BAF remains the same. Also, when tumor sample is not pure, equations (2.1) and (2.2) must account for the impurity by assuming the non-tumor cells have copy number 2 for all genomic locations. Finally, the following equations are used to approximate LRR and BAF:

$$r_i = \gamma \log_2 \frac{2(1-\rho) + \rho(n_{A,i} + n_{B,i})}{\Psi}, \tag{2.3}$$

$$b_i = \frac{1 - \rho + \rho n_{B,i}}{2 - 2\rho + \rho(n_{A,i} + n_{B,i})}, \tag{2.4}$$

where ρ is the tumor cell fraction and γ is a constant (”technology” parameter) depending on the SNP array technology used (0.55 for Illumina). It is also assumed that non-tumor cells have

constant ploidy value (subclonal CNA not alowed). Ψ is the parameter for ploidy. Since for non-tumor cells is $\Psi = 2$, ploidy of the sample can be modeled by $\Psi = 2(1-\rho) + \Psi_t$ where Ψ_t is the tumor ploidy.

By incorporating these variables and assumptions into equations (2.3) and (2.4), we can estimate the parameters n_A and n_B using the following equations:

$$\hat{n}_{A,i} = \frac{\rho - 1 + 2^{\frac{r_i}{\gamma}}(1-b_i)(2(1-\rho)+\Psi_t)}{\rho}, \tag{2.5}$$

$$\hat{n}_{B,i} = \frac{\rho - 1 + 2^{\frac{r_i}{\gamma}}(2(1-\rho)+\Psi_t)}{\rho}. \tag{2.6}$$

The PCF algorithm is used in the segmentation step. We denote by $x_1 < x_2 < ... < x_n$ the probe locations for a given sample and given a genomic location. Let the observed data be (x_i, r_i), and $(x_i, b_i), i = 1, ..., n$. The PCF algorithm aims to minimize the following penalized objective function by dividing a genomic location into segments. This involves partitioning the probes into subsets labeled as $I_1, ...I_Q$, where each subset comprises a series of consecutive probes along the genome. The following criterion is minimized with respect to both the number of segments Q and the assignment of probes to segments:

$$\sum_{j=1}^{Q} \sum_{i \in I_j} [w(r_i - \bar{r}_{s \in I_j})^2 + (1-w)(b_i - \bar{b}_{s \in I_j})^2] + \lambda Q. \tag{2.7}$$

A default value for w in (2.7) is 0.5. The goal is to minimize the number of segments (Q) and the assignment of probes to those segments. The first term in the square brackets in (2.7) represents how well the data fits the Log R values, while the second term represents the goodness

of fit to the BAF data. The criterion also includes a penalty term for discontinuities or change points in the function. A constant (λ) determines the balance between the goodness of fit and the penalty. The input parameters used in this process were a minimum segment length of 6 and $\lambda = 50$.

Grid search algorithm to estimate parameters. Finally, these smoothed data from the segmentation step are used in ASCAT to estimate the parameters ρ (tumor purity), Ψ_t (tumor ploidy), and the absolute allele-specific copy number calls $n_{A,i}$ and $n_{B,i}$. To estimate the parameters ρ and Ψ, a grid-search algorithm is utilized for ρ taking values in $(0.10, 0.11, ..., 1.05)$ and ψ_t taking values in $(1.00, 1.05, ..., 5.40)$. For each pair, the cumulative sum of the calculated total distance to a nonnegative integer solution for the allele-specific copy number profiles across all SNPs in the genome was determined:

$$d(\rho, \Psi_t) = \sum_i w_i \left((\hat{n}_{A,i}(\rho, \Psi_t) - round(\hat{n}_{A,i}(\rho, \Psi_t)))^2 + (\hat{n}_{B,i}(\rho, \Psi_t) - round(\hat{n}_{B,i}(\rho, \Psi_t)))^2 \right), \tag{2.8}$$

where $round$ is used to round the closest non-negative integer. All potential solutions of the data are identified by determining all local minima. For each solution, a goodness-of-fit score is calculated. This score, g, represents a linear rescaling of the total distance to nonnegative integer values into a percentage. Specifically, g is equal to 100% when the distance, denoted as d, is 0, and g is equal to 0 when d is equal to the distance obtained when the allele-specific copy numbers for each SNP differ by 0.25 from nonnegative whole numbers ($\sum_i w_i(2(0.25)^2)$. The value of 0.25 is chosen as a suitable maximum distance, considering that this goodness-of-fit calculation is specifically applied to local minima.

Figure 2.1 gives a visualisation for the results of our CNA analysis using ASCAT for a

patient from TCGA-LIHC cohort. Ploidy is estimated as 2.45, aberrant cell fraction is estimated as 0.43. Red colors represent the minor copy number estimates and green colors represents the major copy number estimates.

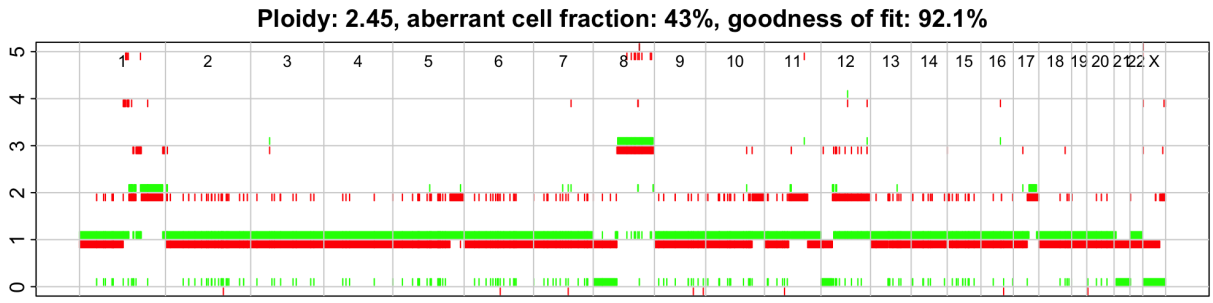


Figure 2.1: Ascat Output

2.2.2 Sequenza algorithm

The Sequenza model (Favero et al., 2015) is a widely used method for detecting CNAs in tumor samples from whole exome sequencing data. This model also assumes the copy number alterations occur at a constant rate along the genome. This assumption implies that the copy number alterations are expected to appear as piecewise constant segments in the copy number profile. In other words, the alterations are assumed to occur in distinct regions rather than being scattered randomly throughout the genome. Furthermore, the tumor purity and ploidy remain constant across the genome. The input data for sequenza consist of two BAM (Binary Alignment Map) files, one derived from the aligned sequencing reads of the tumor specimen and one from the same individual‘s normal specimen. For every position in the genome, the number of reads covering that position, or read depth, is extracted for both the tumor (τ_i) and normal (ν_i) samples. The read depths are then corrected for GC content bias, which is a potential artifact in sequencing that can result uneven coverage of GC content. This bias arises because the proportion of Guanine

(G) and Cytosine (C) nucleotides in a DNA sequence can impact the stability of the DNA strands. Due to the presence of three hydrogen bonds in a G-C pair compared to two in an A-T pair, G-C pairing is more stable. Consequently, DNA strands with higher G-C content have stronger hydrogen bonding, increased stability, and greater resistance to denaturation. Additionally, the read depths are normalized using the overall median read depth to address variability across the genome. Sample data input of Sequenza is given in Table 2.3 which is a combination of tumor and normal BAM files.

chr	position	base.ref	n.depth	t.depth	depth.ratio	Af	Bf	zygosity	GC.percent
chr1	14907	A	8	13	1.62	0.69	0.31	het	58
chr1	16483	G	36	38	1.06	0.97	0	hom	32
chr1	16494	N	27	25	0.95	1	0	hom	30
chr1	16495	G	34	35	1.03	0.63	0.37	het	30
chr1	16534	C	20	21	1.05	0.81	0.19	het	30
chr1	16545	N	10	17	1.82	1	0	hom	56

Table 2.3: Data Input for Sequenza Software

Here n.depth and t.depth are number of reads covering that position in normal (ν_i) and tumor (τ_i) respectively. depth.ratio is unnormalized depth ratio $\frac{\tau_i}{\nu_i}$ zygosity is zygosity of the sample either heterozygous (het) when two bases are detected and the less abundant base accounts for at least 25% of the total reads, or homozygous otherwise. And GC.percent refers to the percentage of guanine (G) and cytosine (C) nucleotides in a DNA sequence.

2.2.2.1 Background of data and model

The objective is to determine the unknown integer values for each genomic position i in the tumor. These values include the copy number n_i, which represents the total number of alleles at position i, and the minor allele copy number m_i, which is the smaller of the two allele-specific

copy numbers. Parameters n_i and m_i are similar parameters as in ASCAT n_A and n_B, minor copy number corresponds to the smaller of these two and the copy number corresponds to the sum of these two. Estimating these parameters involves estimating two real-valued meta-parameters that can also hold biological or clinical significance: the cellularity ρ, which indicates the fraction of tumor cells in the sample, and the ploidy ψ, defined as twice the ratio of tumor DNA mass to normal DNA mass. As in ASCAT, it is assumed that normal genome is constant at all segments and have copy number of 2 for all genomic locations. It is also assumed that tumor cells have constant ploidy value (subclonal CNA not alowed). By assuming that copy numbers tend to remain stable across consecutive genomic positions, we can simplify the representation of n_i and m_i values at each position. This simplification involves grouping them into segments with defined start and end positions. Each segment is then characterized by a copy number value N_s and a minor allele copy number value M_s. Sequenza also utilizes similar sort of PCF algorithm to find the segment boundaries. For each segment, L_s, length of the segment, R_s, mean depth ratio, B_s, mean B allele frequency, S_{R_s}, standard deviation of depth ratio, and S_{R_s}, standard deviation of B allele frequency are calculated.

Probabilistic algorithm to estimate parameters. Finally, these data from segmentation step is used in sequenza to estimate the parameters ρ (tumor cellularity), ψ (tumor ploidy), and the absolute allele-specific copy number calls N_s and M_s using the following probabilistic framework. The likelihood in (2.9) describes the probability of generating the observed depth ratio (R) and B allele frequency (B) measurements for each segment (s) based on specific parameters representing cellularity and ploidy, as well as vectors (N and M) indicating the copy number and number of minor alleles for all segments.

$$p(\mathbf{x}|\theta) = \prod_s p(R_s, B_s|\theta), \tag{2.9}$$

where observed depth ratio and the observed B allele frequency of the segment are assumed to be independent: $p(R_s, B_s|\theta) = p(R_s|\theta)p(B_s|\theta)$. Also, Both R_s and B_s have non-standardized Student's t distribution with degrees of freedom ν is set to 5. This distribution may have been found to provide a good fit to the data or meet certain statistical assumptions required for the analysis. However, further information in the algorithm is not provided to determine the exact basis for this specific choice.

To estimate the parameters, Maximum a posteriori (MAP) estimation is used. It combines prior information or beliefs about the parameter with observed data to obtain an estimate that maximizes the posterior probability. Since the occurrence of copy number 2 is typically more than twice as frequent as any other copy number state, prior probabilities are established for copy numbers (referred to as pNs), with a default higher preference for the solution where the copy number is 2, as shown in Table 2.4.

k	0	1	2	3	4	5	6	7
weight	1	1	2	1	1	1	1	1
pNs=k	0.11	0.11	0.22	0.11	0.11	0.11	0.11	0.11

Table 2.4: Sequenza Prior Probabilities

To estimate meta parameters ρ and ψ, grid based search is utilized with ρ ranging from 0.1 to 1 in steps of 0.01, and ψ ranging from 1 to 7 in steps of 0.1. Figure 2.2 gives a visualisation for the results of our CNA analysis using Sequenza for the same patient in Figure 2.1. Ploidy is estimated as 2.2, aberrant cell fraction is estimated as 0.44. Blue colors represent the minor copy

number estimates and red colors represents the major copy number estimates.

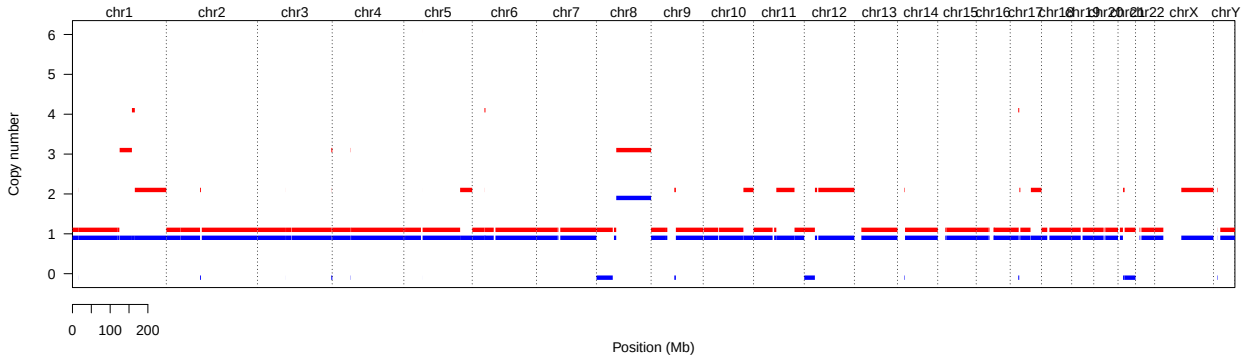


Figure 2.2: Sequenza Output

2.3 Circular Binary Segmentation (CBS) algorithms

Circular Binary Segmentation (CBS) algorithm (Olshen et al., 2004) is another method of analysis of copy number alterations in tumor genomes. The key components and steps involved in CBS-based algorithms for CNA analysis are as follows:

1. Data preprocessing: The input data, such as DNA sequencing or microarray data, is preprocessed to remove noise, correct biases, and normalize the copy number measurements. This step ensures the data is in a suitable format for subsequent analysis.

2. Segmentation: The core step of CBS algorithms is the segmentation of the genome into regions with similar copy number profiles. This is achieved by iteratively dividing the genome into segments and testing the statistical significance of the differences in copy number between adjacent segments. The segmentation is performed in a circular manner, ensuring that the algorithm captures large-scale genomic events that may span the ends of chromosomes.

3. Binary segmentation: The binary segmentation approach divides the genome into two

segments and tests the statistical hypothesis of equal copy number between them. If the hypothesis is rejected, indicating a significant difference in copy number, the segments are further divided. This process continues recursively until no further significant differences are detected or until a predefined stopping criterion is met.

4. Statistical testing: CBS algorithms employ statistical tests, such as t-tests or permutation tests, to assess the significance of copy number differences between adjacent segments. The choice of the statistical test depends on the nature of the data and assumptions about its distribution.

5. Merge and refine segments: After the segmentation process, neighboring segments with similar copy number profiles are merged to form larger regions. This helps to identify broader regions of copy number alterations and reduce the impact of noise and spurious breakpoints. Additional refinement steps, such as outlier detection or smoothing, may be applied to improve the accuracy and robustness of the segment boundaries.

6. Copy number estimation: The inferred segments represent regions of the genome with similar copy number profiles. The copy number state for each segment is estimated based on the average copy number within the segment. This provides a quantitative assessment of the copy number alterations in the tumor genome.

7. Visualization and interpretation: The resulting segmented copy number profile can be visualized using plots or heatmaps, highlighting regions of copy number alterations. Researchers can analyze and interpret the identified segments to gain insights into the genomic events associated with the tumor, such as amplifications, deletions, or structural rearrangements.

Next, we discuss FACETS algorithm that utilizes CBS in the segmentation step of copy number estimation.

2.3.1 FACETS algorithm

FACETS (Shen and Seshan, 2016) is another package designed for the analysis of copy number alteration using sequencing data. It is applicable to various sequencing platforms, including whole-genome, whole-exome, and targeted cancer gene panels. FACETS offers a comprehensive analysis pipeline that encompasses various steps for processing BAM (Binary Alignment Map) files including library size normalization and GC-normalization, joint segmentation of both total and allele-specific signals, allowing for accurate identification of copy number alterations. To reduce hypersegmentation in regions of the genome with dense single nucleotide polymorphisms (SNPs), subsampling is performed within 150-250 base pair intervals. This helps to ensure that the analysis focuses on larger genomic segments rather than individual SNPs. For every genomic position, logR is determined as the logarithm of the ratio between the total read depth observed in the tumor sample and the corresponding depth in the normal sample. On the other hand, logOR is calculated as the logarithm of the odds ratio between the variant allele count detected in the tumor sample and the count in the normal sample. Sample data input of FACETS is given in Table 2.5.

Chr	Position	File1R	File1A	File1E	File1D	File2R	File2A	File2E	File2D
1	69424	170	117	0	0	158	103	0	0
2	69515	0	76	0	0	0	77	0	0
3	69536	103	0	0	0	99	0	0	0
4	808866	96	0	0	0	133	0	0	0
5	809120	66	0	0	0	105	0	0	0
6	809176	79	0	0	0	126	0	0	0

Table 2.5: Data Input for FACETS Algorithm

Here, File1R, File1A, File1E and File1D columns are the counts of number of reads with the

reference allele, alternate (variant) allele, errors (neither ref nor alt) and deletions in that position for normal sample and File2R, File2A, File2E and File2D are the same counts for tumor sample.

2.3.1.1 Background of data and model

Let Φ denote the variant tumor cellularity which is a function of tumor purity, ρ and clonal frequency. Also denote (m, p) as the parental copy numbers similar to the n_A and n_B parameters in section 2.2.1. The expected $logR$ and $logOR$ are:

$$E[logR] = \log\left(\frac{(m+p)\Phi}{2} + (1-\Phi)\right) + w(\cdot) + \lambda, \tag{2.10}$$

$$E[logOR] = \log\left(\frac{m\Phi + 1 - \Phi}{p\Phi + 1 - \Phi}\right) \tag{2.11}$$

where $w(\cdot)$ accounts for the systematic bias. Since $logR$ quantifies relative copy number, λ is a constant responsible for the conversion to absolute copy number. FACETS utilizes an extended CBS algorithm for change point detection where a bivariate genome segmentation is performed on $logR$ and $logOR$ values using a bivarite Hotelling T^2 statistic (Shen and Seshan, 2016). If the maximum statistic calculated surpasses a predetermined critical value (cval), it signifies the presence of a change, and the change points that yield the highest statistic are identified. Following the segmentation process, a clustering algorithm is employed to group the segments together based on their shared underlying genotype.

Probabilistic algorithm to estimate parameters. Finally, segmentation data is used in FACETS to estimate the cellular fraction and integer copy numbers (major and minor). This is achieved by modeling the expected values of $logR$ and $logOR$, given the total (t) and parental (m,

p) copy numbers, as a function of a parameter. A combination of parametric and non-parametric methods is utilized to achieve this, enabling the modeling of both clonal and subclonal events.

The expectation-maximization (EM) algorithm is used to maximize the likelihood of the joint data by treating it as an estimation problem, where the hidden or latent copy number states are considered as ”missing” data. Figure 2.3 gives a visualisation for the results of our CNA analysis using FACETS for the same patient in Figures 2.1 and 2.2. Ploidy is estimated as 2.2, aberrant cell fraction is estimated as 0.5. Red colors represent the minor copy number estimates and black colors represents the major copy number estimates.

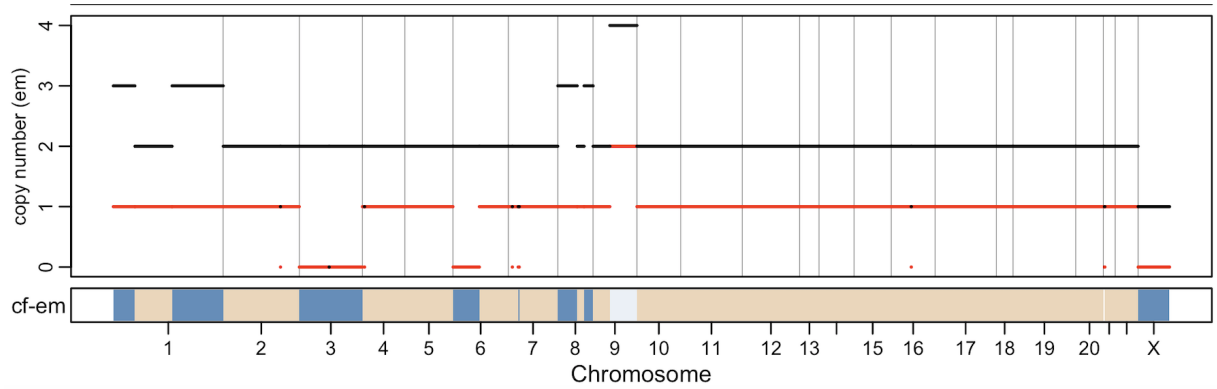


Figure 2.3: FACETS Output

2.4 Hidden Markov Model (HMM) based algorithms

Hidden Markov Model (HMM) based algorithms (Fridlyand et al., 2004) are widely used in the analysis of copy number alterations (CNAs) in tumor genomes. HMM-based algorithms utilize the probabilistic framework of HMMs to model the underlying copy number states and infer the most likely copy number profile of a tumor sample.

The key components and steps involved in HMM-based algorithms for CNA analysis are as follows:

State representation: The copy number states in the genome are represented as hidden states in the HMM. Each hidden state corresponds to a specific copy number level or state, such as deletion, normal copy number, or amplification. The number of hidden states is determined based on the expected range of copy number alterations in the tumor genome.

Observations: The observed data, such as DNA sequencing or microarray data, are associated with each hidden state in the HMM. The observations can be continuous or discrete, representing the measurements of copy number at specific genomic loci. For example, in microarray data, the fluorescence intensity ratios can be used as observations.

Transition probabilities: HMMs incorporate transition probabilities that define the likelihood of transitioning from one copy number state to another. These transition probabilities capture the temporal and spatial dependencies between adjacent genomic loci. They can be estimated from the training data or set based on prior knowledge of the genomic structure and the expected patterns of CNAs.

Emission probabilities: The emission probabilities model the likelihood of observing a particular data point (copy number measurement) given the underlying hidden state. Emission probabilities can be estimated using statistical methods, such as maximum likelihood estimation or Bayesian inference. The choice of the emission probability distribution depends on the nature of the observed data (e.g., Gaussian distribution for continuous data or multinomial distribution for discrete data).

Model training: HMM-based algorithms involve training the model parameters, including the transition probabilities and emission probabilities, based on the observed data. This typically involves an iterative process, such as the Baum-Welch algorithm or expectation-maximization (EM) algorithm, to maximize the likelihood of the observed data given the model.

Inference and decoding: Once the HMM is trained, it can be used to infer the most likely copy number profile for a given tumor sample. The Viterbi algorithm or forward-backward algorithm is commonly used for decoding the most probable sequence of hidden states (copy number states) that generated the observed data. This provides the inferred copy number alterations in the tumor genome.

Post-processing and interpretation: After decoding the copy number profile, post-processing steps may be applied to refine the results and improve their biological interpretability. This can include noise reduction, smoothing, outlier detection, and merging or splitting of regions based on certain criteria. The resulting copy number profile can be visualized and analyzed to identify significant CNAs and their boundaries, potentially revealing important genomic alterations associated with the tumor.

The final copy number algorithm we discuss next utilizes Hidden Markov Models (HMMs) to detect change points in the genome.

2.4.1 hsegHMM algorithm

The hsegHMM model (Choo-Wosoba et al., 2018) is a novel approach for detecting copy number alterations in tumor samples using sequencing data. Unlike traditional methods, hsegHMM takes into account the phenomenon of hypersegmentation, where short segments with insignificant copy number changes can introduce noise and fragmentation in the analysis. The model is based on a hidden Markov model (HMM) framework and utilizes an efficient expectation-maximization (E-M) algorithm to infer copy number profiles. Similar to previous algorithms, two input files are required, one containing the sequencing reads aligned to the tumor specimen and another

containing reads aligned to the matched normal specimen. Sample data input of hsegHMM is the same as in Table 2.5.

Background of data and model. The variant and reference read counts used in this study were extracted from matched tumor-normal BAM files. These files contained genomic sequencing data from both the tumor specimen and the corresponding normal sample. The read counts were specifically obtained for germ-line SNP locations, which were compiled from reference libraries derived from various healthy tissues. At each SNP position, $logR$ is defined as the log-ratio of total read depth in the tumor versus that in the normal and $logOR$ is defined by the log-odds ratio of the variant allele count in the tumor versus in the normal. Let Z_k be the unknown genotype as the hidden state of the kth genomic location and Y_k and X_k be the corresponding $logR$ and $logOR$ respectively. 12 different states of Z_k specified in Choo-Wosoba et al. (2018) are shown in Table 2.6.

State (j)	Genotype	Copy number (CT)	Allelic information
1	0	0	HOMD
2	A	1	DLOH
3	AA	2	NLOH
4	AB	2	HET
5	AAB	3	GAIN
6	AAA	3	ALOH
7	AAAB	4	ASCNA
8	AABB	4	BCNA
9	AAAA	4	ALOH
10	AAAAB	5	ASCNA
11	AAABB	5	UBCNA
12	AAAAA	5	ALOH

Table 2.6: hsegHMM HMM States

The abbreviations in Table 2.6 are homozygous deletion (HOMD), hemizygous deletion LOH (DLOH), copy neutral LOH (NLOH), diploid heterozygous (HET), gain of 1 allele (GAIN),

amplified LOH (ALOH), allele-specific copy number amplification (ASCNA), balanced copy number amplification (BCNA), and unbalanced copy number amplification (UBCNA).

Denote $j = 1, .., J$ as the index for hidden states and $k = 1, ..., N$ as the index for genomic position. Then, the expectations of $logR$ and $logOR$ given the state space are

$$E(Y_k|Z_k = j) = \mu_j = log\left[\frac{(1-\rho)C_N + \rho C_{T,j}}{\psi}\right], \tag{2.12}$$

and

$$E(X_k|Z_k = j) = \zeta_j = log\left[\frac{(1-\rho) + \rho m_j}{(1-\rho) + \rho p_j}\right], \tag{2.13}$$

where m_j and p_j are the maternal and paternal copy numbers of the tumor at the kth genomic location respectively and ρ is tumor purity and ψ is ploidy. C_N is the copy number of normal cells (default 2) and $C_{T,j}$ denotes the copy number of tumor cells at the j^{th} state. The model assumptions are:

1. $\mathbf{Z} = \{Z_1, Z_2, ..., Z_N\}$, the genotype sequence across chromosomes follows a Markov chain with transition probabilities $P_{ij} = P(Z_k = j|Z_{k-1} = i)$ and an initial probability, $r_{0j} = P(Z_1 = j)$.

2. Y_k follows a t distribution with degree of freedom ν that is a mixture of normal distribution with a gamma distribution:

$$t_\nu(Y_k|Z_k = j) = \int_{u_k} \mathcal{N}\left(Y_k|\mu_j, \frac{\kappa^2}{u_k}\right) G\left(u_k, \frac{v}{2}, \frac{v}{2}\right) du_k,$$

where κ^2 does not vary with j.

$$\kappa_j^2 = (1 + \sqrt{(1 - M_j)^2 + (1 - m_j)^2})\kappa^2.$$

For normal cancer cells $M_j = m_j = 1$

3. X_k^2 follows a non-central chi-square distribution with degree of freedom 1 and non-centrality parameter $\delta_j = \frac{\zeta_j^2}{\tau^2}$:

4. Y_k and X_k are conditionally independent given the genotype state.

HMM parameters P_{ij} and r_{0j} under constraints $\sum_{j=1}^{J} P_{ij} = 1$ and $\sum_{j=1}^{J} r_{0j} = 1$, and global parameters $\theta = \{\alpha, \psi, \kappa^2, \tau^2, \nu\}$ are estimated using E-M algorithm.

2.5 Comparative analysis of copy number detection tools

We conduct a comparative analysis between the ASCAT method and other commonly used methods, including Facets, Sequenza, and HsegHMM, for copy number alteration (CNA) detection in cancer genomics. The motivation behind this comparison arises from the availability of different data sources for these methods on the TCGA-LIHC cohort. ASCAT relies on SNP-array data, whereas Facets, Sequenza, and HsegHMM primarily utilize whole-exome (WEX) sequencing data.

In our study, we encounter a scenario where SNP-array data was only available for one cohort, while WEX data was available for all three liver cancer cohorts. This discrepancy in data availability prompted us to validate the performance of ASCAT and investigate its suitability as a validation tool in the absence of SNP-array data.

Accurate detection and characterization of CNAs are crucial for comprehending the genomic

landscape of cancer, including its impact on disease progression and treatment response. Each computational method for CNA detection possesses its own strengths and limitations. ASCAT provides high-resolution copy number estimates. On the other hand, Facets, Sequenza, and Hseg primarily rely on WEX data, which is more readily accessible across diverse cancer cohorts.

We acknowledge the limitations in data availability, with SNP array data being available for only one cohort. Consequently, we predominantly utilized Sequenza as the primary method in our research, given its compatibility with WEX data, which was available for the majority of our cohorts. This choice exemplifies the practical considerations surrounding data availability and cost-effectiveness.

In conclusion, by comparing ASCAT's results with those obtained from WEX-based methods, we contribute to the understanding of CNA detection methodologies and offer guidance to researchers in selecting appropriate approaches based on data availability and research objectives.

2.5.1 Metric to measure differences between tools

We define a measure to quantify the differences in major and minor copy number estimates between two different sources. Let M_{1i} and M_{2i} be the major copy number estimates, and m_{1i} and m_{2i} be the minor copy number estimates of the i-th genomic location of a sample from software 1 and software 2, respectively. We calculate the differences d_M and d_m for each individual sample as follows:

$$d_M = \frac{1}{l_G} \sum_{j}^{22} \sum_{i=1}^{l_j} \frac{|M_{1i} - M_{2i}|}{l_j}, \tag{2.14}$$

$$d_m = \frac{1}{l_G} \sum_{j}^{22} \sum_{i=1}^{l_j} \frac{|m_{1i} - m_{2i}|}{l_j}, \tag{2.15}$$

where l_j is the length of chromosome j and l_G is the length of the whole genome. If they are close to zero, we say that tools agree each other well. Finally, the dissimilarity measure, we use to compare two tools is defined as

$$d = \frac{d_m + d_M}{2}, \tag{2.16}$$

where $d \geq 0$.

We provide an overview of the distribution of patients across different difference intervals in Table 2.7. Determining an acceptable agreement range for ASCAT and Sequenza measurements can be challenging. However, upon analysis, we observe that the majority of the differences fall below 0.5. This observation suggests a level of agreement between these tools, indicating concordance in their estimations.

d measurements	**Number of Patients**
0.0 - 0.1	186
0.1 - 0.2	66
0.2 - 0.5	23
0.5 - 0.9	20
0.9 - 1	23
≥ 1	35

Table 2.7: Number of Patients per Difference Interval

2.5.2 Purity and ploidy estimates

In our study, we focus solely on comparing ASCAT with the tool Sequenza across the genome. This decision was motivated by the need to avoid the time-consuming and resource-

intensive process of reference genome conversion for each dataset. During our analysis, we utilized read count data (WEX data) aligned to the GRCh38 genome version. Sequenza, FACETS, and HsegHMM were applied to the GRCh38 genome version, while ASCAT results are for the GRCh37 reference genome from SNP-array experiments. This discrepancy necessitated either converting the ASCAT results to the GRCh38 reference genome or performing a new alignment of the WEX data to GRCh37. However, this conversion of reference genome coordinates for ASCAT presented challenges when comparing position information between the two most recent builds of the human genome. It is important to note that these genome builds typically have non-comparable position information (Ormond et al., 2021).

Given the time and resource constraints of our study, we focused on comparing ASCAT and Sequenza. We already had Sequenza results obtained using the GRCh37 reference genome with alignment conversion performed using the CrossMap tool (Zhao et al., 2014). Prioritizing this comparison allowed for a direct assessment of performance between ASCAT and Sequenza without the additional complexities and time associated with reference genome conversion. This approach facilitated an efficient evaluation of the similarities and differences between ASCAT and Sequenza in our dataset. Although we did not directly compare ASCAT with other tools across the entire genome, it is important to highlight that we still obtained ploidy and purity estimates from each tool for every sample. These estimates provide information about the overall genomic characteristics and tumor purity in each sample. Ploidy, which refers to the number of chromosome sets in a cell, plays a crucial role in understanding the genomic complexity and stability of tumors. Estimating ploidy using various computational tools provides information on overall chromosomal content and facilitates the identification of aneuploidies and other large-scale genomic alterations. In Figure 2.4, we present Kaplan-Meier survival curves for patient

groups categorized as ”High” or ”Low” based on median ploidy values estimated from four different tools: ASCAT, Facets, Sequenza, and hsegHMM. Our analysis reveals that ASCAT shows the most significant separation in survival outcomes.

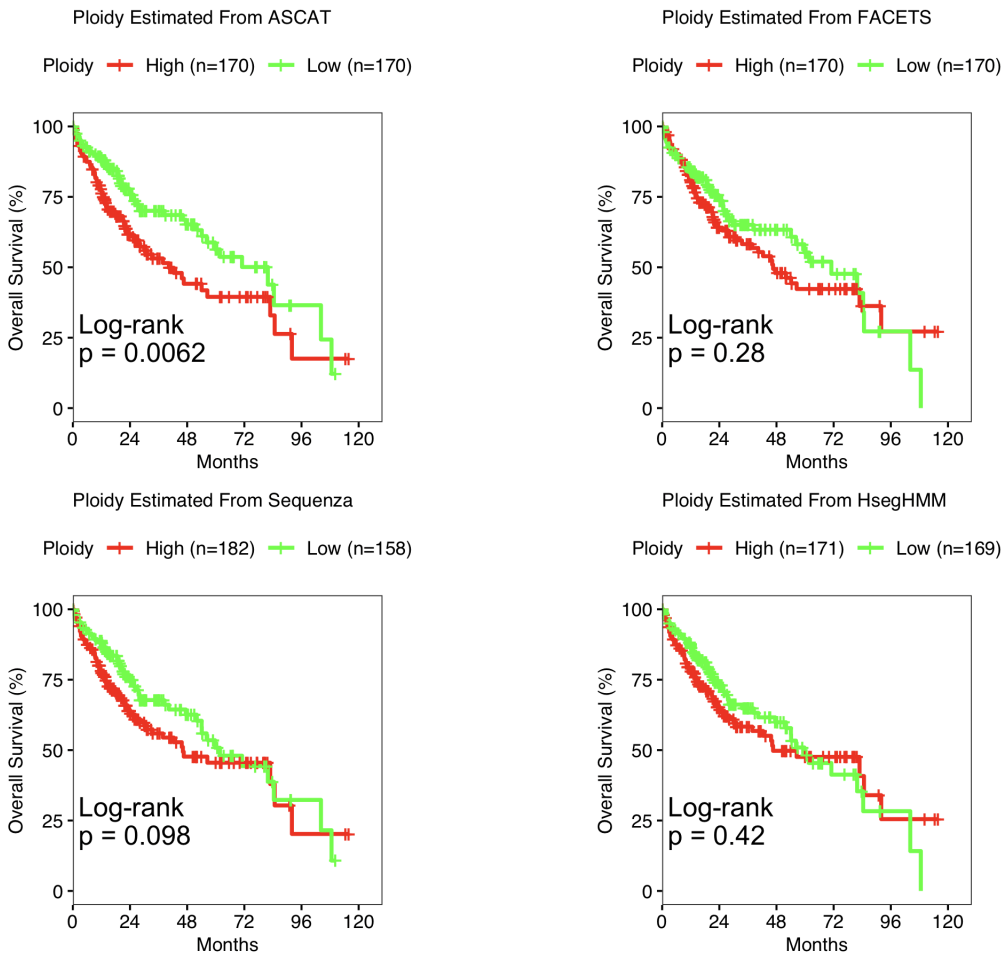


Figure 2.4: Survival curves of patient groups classified as ”High” or ”Low” ploidy values based on median values estimated from 4 different tools.

Considering that higher ploidy levels indicate abnormalities in the cancer genome, we suggest that the results obtained from ASCAT may have stronger biological validity. However, due to the unavailability of the required input data for ASCAT in the two additional cohorts, an alternative tool needs to be employed for the analysis. In this context, it is wise to select a tool that aligns well with ASCAT to ensure consistency and biological validity in the obtained

results. By choosing a tool that agrees with ASCAT, we can enhance the reliability and biological interpretability of our findings.

Purity, on the other hand, represents the proportion of tumor cells in a sample compared to the total number of cells, including normal cells. Estimating tumor purity is crucial for interpreting copy number alterations and identifying the genomic regions that are specifically affected by the tumor. The purity estimates obtained from each tool allow us to assess the tumor cellularity and determine the extent to which the observed copy number alterations are tumor-specific.

In Figure 2.5, we present the Bland-Altman analysis, which compares the ploidy and purity estimates obtained from different tools. The differences between the measurements (y-axis) are calculated as the values obtained from ASCAT minus the measurements from the other tools. The averages (x-axis) are the average values between ASCAT and the respective tool.

This analysis allows us to assess the agreement between ASCAT and the other tools by examining the distribution of the differences. If the estimates from ASCAT and the other tools are in close agreement, we would expect the differences to be centered around zero, indicating minimal bias. However, if there are systematic differences or biases between the measurements, we would observe deviations from zero.

In the Bland-Altman plot, the red points represent individuals whose measurements differ more than two standard deviations from the mean difference. These points highlight cases of significant disagreement between ASCAT and the other tools, suggesting potential inconsistencies or discrepancies in their estimations.

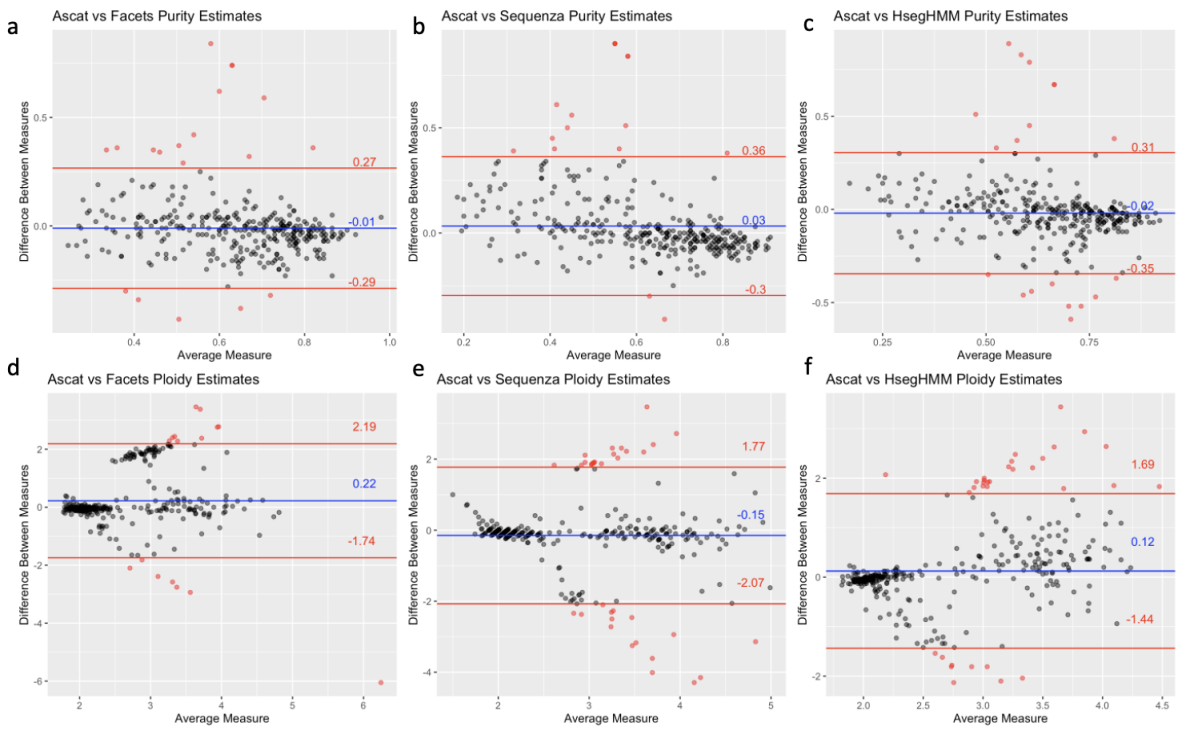


Figure 2.5: Bland Altman Plots: Each dot represents an individual comparison between ASCAT estimates and estimates from another tool (a,d) Facets, (b,e) Seqeunza and (c,f) HsegHMM. The x-axis represents the average value of the measurements, while the y-axis represents the difference between ASCAT estimate and the estimate from the other tool. The first row consists of three plots comparing purity the measurements, and the second row consists of three plots comparing the ploidy measurements.

In Tables 2.8 and 2.9, the mean difference represents the average deviation between the measurements obtained from ASCAT and the alternative tools. In the case of ASCAT vs. Sequenza, the mean difference for purity is 0.03, indicating a slight positive deviation. Similarly, for ASCAT vs. Facets and ASCAT vs. HsegHMM, the mean differences are -0.01 and -0.02, respectively, suggesting small negative deviations. The limits of agreement (LoA) represent the range within which the differences between the methods are expected to fall. In the ASCAT vs. Sequenza comparison, the LoA range for purity is 0.66, indicating that most of the differences between the two methods lie within this range. For ASCAT vs. Facets and ASCAT vs. HsegHMM, the

LoA ranges are 0.55 and 0.65, respectively. In the Bland-Altman plot, the red points represent individuals whose measurements differ more than two standard deviations from the mean difference. These points highlight cases of significant disagreement between ASCAT and the alternative tools, suggesting potential inconsistencies or discrepancies in their estimations.

Based on the results, we observe a relatively good agreement between ASCAT and the other tools in terms of purity and ploidy estimation. While there are slight differences in the mean difference and limits of agreement, the overall agreement is promising. Considering the popularity, ease of usage, and the favorable agreement results, we have made the decision to utilize Sequenza as an alternative to ASCAT for copy number estimation. Sequenza demonstrates a high level of concordance with ASCAT, making it a viable and practical choice for our analysis. It is important to acknowledge that no method is perfect, and each tool may have its strengths and limitations. However, the strong agreement observed between ASCAT and Sequenza provides confidence in the reliability and validity of Sequenza for our specific analysis needs. Further assessments and validations can be conducted to fully evaluate the performance of Sequenza in comparison to ASCAT and other tools.

Comparison	**Mean Difference**	**Limits of Agreement Range**
ASCAT vs. Sequenza	0.03	0.66
ASCAT vs. Facets	-0.01	0.55
ASCAT vs. HsegHMM	-0.02	0.65

Table 2.8: Comparison of Methods for Purity Estimates

Comparison	**Mean Difference**	**Limits of Agreement Range**
ASCAT vs. Sequenza	-0.15	3.84
ASCAT vs. Facets	0.22	3.93
ASCAT vs. HsegHMM	0.125	3.13

Table 2.9: Comparison of Methods for Ploidy Estimates

2.6 Discussion

In this chapter, we provided an overview of copy number alteration (CNA) detection tools and focused on the use of whole exome sequencing (WEX) data for CNA identification and analysis in tumor cells. We also briefly mentioned the utilization of SNP array data for comparison purposes.

We reviewed and compared four CNA detection tools: ASCAT, Sequenza, FACETS, and hsegHMM. These tools employ different methods and algorithms for CNA detection, such as Piecewise Constant Fitting (PCF) algorithms, Circular Binary Segmentation (CBS) algorithms, and Hidden Markov Model (HMM)-based algorithms. PCF algorithms approximate the copy number profile of a tumor sample as a piecewise constant function and segment the genome into regions with distinct copy number states. ASCAT and Sequenza are examples of tools utilizing PCF algorithms. ASCAT estimates allele-specific copy number profiles from SNP array data using a Bayesian hierarchical model, while Sequenza infers tumor purity, ploidy, and copy number profiles from tumor-normal paired WEX data using statistical algorithms and binomial mixture models.

FACETS is a computational tool commonly used in DNA sequencing data analysis for estimating tumor purity, ploidy, and allele-specific copy number profiles. It employs a likelihood-based model and segmented read counts to accurately estimate these genomic characteristics. HsegHMM, on the other hand, builds upon the initial steps of the FACETS algorithm but incorporates hidden Markov models (HMMs) for change point detection. By utilizing the Viterbi algorithm and a maximum likelihood approach, hsegHMM identifies genomic regions with altered copy numbers and provides detailed insights into the allele-specific changes within the genome.

Comparative analysis of these tools provides information about their underlying models and usages. The choice of tool depends on various factors, including the data requirements and methodological approaches.

Data Requirements: ASCAT primarily utilizes SNP array data, while Sequenza, FACETS, and hsegHMM analyze WEX data. Researchers should consider the available data types when selecting a tool.

Methodological Approaches: ASCAT employs a Bayesian hierarchical model, while Sequenza, FACETS, and hsegHMM utilize statistical algorithms and models like binomial mixture models, likelihood-based models, and Hidden Markov Models, respectively. Researchers should assess the suitability of these methodologies for their research objectives and data characteristics.

It is important to note that the comparison of these tools can be challenging due to the lack of a gold standard for performance evaluation. In this thesis, we compared them on the liver cancer data. Each tool has its strengths and limitations, and the choice of tool should align with specific requirements, data characteristics, and research objectives.

The field of CNA detection tools is rapidly evolving, with continuous development of new methods and algorithms. Staying updated with the latest advancements and selecting the most appropriate tool for a specific study is essential. Future research should focus on benchmarking and evaluating the performance of CNA detection tools using standardized datasets to facilitate accurate and reliable CNA analysis in cancer research.

In conclusion, CNA detection is a vital aspect of cancer genomics research, and the selection of appropriate tools and methods is crucial for accurate and reliable clonality analysis. This chapter provided an overview of CNA detection tools, compared their features, and discussed the main steps involved in the algorithms that those tools utilize. Researchers should consider the

strengths and limitations of each tool and tailor the analysis approach to the specific characteristics of the data. Further research and development in this field will contribute to a more accurate and comprehensive understanding of CNAs and their role in cancer progression and treatment.

Chapter 3: Clonality and clustering analysis of liver cancer data

3.1 Overview

Each patient's tumor in liver cancer, particularly HCC (Hepatocellular Carcinoma), is composed of different proportions of subclones resulting from tumor cell evolution. In order to unravel the ancestral relationship among these clones, we employ phylogenetic tree estimation on tumor samples using mutation read-count data. Our focus is on functional mutations occurring in genes related to the liver or liver cancer.

To accomplish this, we utilize the SMASH (Subclone Multiplicity Allocation and Somatic Heterogeneity) approach, a frequentist method (Little et al., 2019). SMASH clusters somatic mutations, taking into account their corresponding copy number alteration estimates, to identify subclones within tumor samples. We explore various tree topology configurations encompassing 1-5 subclones, and for each configuration, we compute the maximum likelihood for the SMASH model parameters. This process enables us to infer the most probable tree configuration that best represents the evolutionary relationships among the subclones. In the subsequent section, we will provide a detailed description of the SMASH model, including the underlying assumptions and constraints. Additionally, we will discuss the probabilistic framework within which all tree models, subject to certain constraints, are considered equally probable.

Figure 1.1 provides a visual summary of the analysis conducted in this Chapter. The figure

is divided into three panels, each representing a key aspect of the research.

Clonality Analysis: The first panel focuses on the clonality analysis. In (a), we describe how SMASH takes mutation read count data (WEX) as input and generates both linear and nonlinear phylogenetic trees. To refine our analysis, we specifically select functional gene mutations for the clonality analysis. In (b), we present the overall survival of patients based on the linear and nonlinear phylogenies for the TCGA-LIHC cohort. We examine survival outcomes separately for functional mutations (on the left) and all mutations (on the right). We also provide the survival curves for linear versus nonlinear trees of TIGER-LC cohorts in Figure 3.7 B.

Clustering Analysis: The second panel summarizes the clustering analysis performed on the nonlinear phylogenetic trees. In (c), we describe the creation of features using modified probability vectors obtained from SMASH outputs. These features capture essential characteristics of subclones within the nonlinear phylogenetic trees. In (d), we detail the implementation of a clustering algorithm that utilizes the created features and the nonlinear trees to identify shallow branching and deep branching tree groups. To ensure the reliability of the clustering results, stability analysis is conducted in (e) to confirm the presence of well-separated and stable clusters.

Implications for Clinical Outcomes: The final panel provides an overview of the implications of the research findings on clinical outcomes. We explore the potential implications of our clonality and clustering analysis results in terms of patient prognosis, identification of potential drivers of tumor evolution, and the characterization of the TME (Tumor Micro-environment) across distinct tumor phylogenies. By investigating the clonal architecture and evolutionary patterns of liver cancer through phylogenetic tree reconstruction and clustering analysis, this chapter enhances our understanding of tumor heterogeneity, subclone dynamics, and their impact on clinical outcomes in Hepatocellular Carcinoma.

3.2 SMASH: Subclone multiplicity allocation and somatic heterogeneity

SMASH (Little et al., 2019) is a statistical tool to identify clonal structure of tumor using mutation read counts after correcting for copy number alterations (CNA). All possible trees that are compatible with the observed read counts and observed CNA (estimated by another tool) are enumerated and the probability of each tree is quantified according to the model described in this Section. For each patient, we have reference and alternate read counts of a set of mutations on specific locations. Table 3.1 gives a sample input file for the SMASH algorithm, where Ref Counts refers to the number of reads supporting the reference allele and Var Counts refers to the number of reads supporting the mutated allele. Also, Minor CNA and Major CNA are estimated copy numbers from one of the external software such as the ones discussed in Chapter 2 either using WEX (Figure 1.2) or SNP-arrays (Figure 1.4).

Mutation ID	**Ref Counts**	**Var Counts**	**Normal CN**	**Minor CN**	**Major CN**
ATAD3B_chr1_1495680	94	6	2	1	1
KIF1B_chr1_10365418	127	28	2	1	1
PDE4DIP_chr1_149026742	4	2	4	1	3
FCRL1_chr1_157796103	181	18	5	1	4
NR1I3_chr1_161231141	149	23	5	1	4
KIF26B_chr1_245611871	110	20	3	1	2
NCK2_chr2_105881477	64	3	2	1	1
BSN_chr3_49662163	79	21	2	1	1
BSN_chr3_49663323	62	15	2	1	1
ROBO1_chr3_78657166	105	34	2	1	1

Table 3.1: Data input for SMASH

The following assumptions are made in Little et al. (2019) to construct a clonal tree of a tumor sample.

1. Primary tumors arise from a founder clone or have unicellular origin.

2. Loci harboring mutations have homozygous reference alleles in normal cells and a mixture of reference and alternate alleles in tumor.

3. Each mutation event occurs only once on a single allele and a locus will not undergo more than one point mutation or revert back to its original base (infinite site assumption).

4. At most two descendant subclones can evolve from an ancestral subclone.

5. Copy number alterations in tumors are clonal, indicating that all tumor cells possess the same estimated copy number profile.

Assumption (1) is derived from the clonal evolution theory of tumor growth. Assumption (2) is automatically met because genetic loci with germline mutations are eliminated during somatic mutation calling. Assumption (3) is known as the infinite site assumption (Hudson, 1983), which is reasonable given the small number of mutated loci compared to the genome's size. Assumption (4) is sensible when considering tumor evolution in a more precise time frame and helps reduce the number of possible phylogenies to consider. Assumption (5) is particularly relevant to the copy number inference method we employ, Sequenza. It imposes constraints on the copy number evolution patterns and aids in reducing the number of potential phylogenies to be explored.

Observed data: Let $l = 1, ..., L$ index each locus harboring a mutation (after mutation calling and filtering), and denote alternative and reference read counts from tumor sample by A_l and R_l respectively. Also, let (C_{l1}, C_{l2}) be clonal copy number for each locus harboring a mutation representing major and minor copy number states respectively.

Tree enumeration: Assume there are S subclones for the tumor sample each relating each other through a phylogenetic tree. We denote possible allocation of somatic mutations by a vector of

length S where each element is indicator of whether this mutation occurs in the s^{th} subclone: $q_u^T = (q_{u1}, ...q_{uS})$.

Let Q_k represent the set of allocations for the k^{th} enumerated tree. There are 13 enumerated trees for $S = 1 : 5$ subclones considering clonal tree assumptions. If a clonal mutation occurs in a copy number altered region, the multiplicity of it is inferred either 1 (CNA occurs before mutation) or one of C_{l1} and C_{l2} (mutation occurs before CNA). If a subclonal mutation occurs in a copy number altered region, its multiplicity will be always 1 due to the clonal CNA assumption. Thus, Possible multiplicities are

$$M_l = \{m \mid m > 0, m \in unique(1, C_{l1}, C_{l2})\},$$

where $unique(Z)$ are the unique elements of set Z.

Subclone proportions: Assume tumor sample has S subclones, and η_s denotes the proportion of cells that belong to subclone s. Then, tumor purity is given $\phi = \sum_{s=1}^{S} \eta_S$. Also, subclone proportions in the cancer cell population are denoted by $\vartheta^{\mathbf{T}} = (v_1, ..., v_s)$ where $v_s = \frac{\eta_s}{\phi}$.

Given a tree structure and copy number estimates, A_l given $T_l = A_l + R_l$ is modeled by a mixture of binomial distributions. Tumor purity and copy number states are given (estimated by Sequenza). Suppose there are W unique copy number states $(c_1, ..., c_W)$ and given c_w there are D_w possible combinations of allocation and multiplicity. Denote the d^{th} combination by $e_{wd} = (q_d, m_{wd})$ where q_d represents allocation (depends on tree structure not copy number) and m_{wd} represents multiplicity (depends on copy number). Let U_l and M_l be latent random variables for the allocation and multiplicity of lth mutation respectively. For each D_w combination, denote $\pi_{\mathbf{w}} = (\pi_{w1}, ..., \pi_{wD_w})$ as mixture proportions. Assume the number of variants unexplained by

combinations of U_l and M_l follows a discrete uniform distribution with proportion parameter denoted by ϵ. Denote $\Theta = (\epsilon, \vartheta, \pi_{\mathbf{w}})$.

Binomial mixtures: Let $E_l = (U_l, M_l)$ and $G_l = (T_l, C_l, \phi, \Theta)$. For l^{th} mutation,

$$P(A_l \mid G_l, C_l = c_w) = \epsilon \frac{1}{T_l} + (1-\epsilon) \sum_{d=1}^{D_w} P(E_l = e_{wd}, A_l \mid G_l, C_l = c_w)$$

$$= \epsilon \frac{1}{T_l} + (1-\epsilon) \sum_{d=1}^{D_w} P(E_l = e_{wd} \mid G_l, C_l = c_w) P(A_l \mid E_l = e_{wd}, G_l, C_l = c_w)$$

$$= \epsilon \frac{1}{T_l} + (1-\epsilon) \sum_{d=1}^{D_w} \pi_{wd} P(A_l \mid E_l = e_{wd}, G_l, C_l = c_w).$$

Then, the likelihood for L mutations is proportional to

$$\prod_{w=1}^{W} \prod_{l: C_l = c_w} P(A_l \mid G_l, C_l l = c_w), \tag{3.1}$$

where $A_l \mid E_l = e_{wd}, G_l$ has Binomial distribution with $n = T_l$ and

$$p = p_{wd} = \frac{m_{wd} \phi \vartheta^T q_d}{(C_{l1} + C_{l2})\phi + 2(1-\phi)}.$$

Parameter estimation: The process of maximizing the likelihood in (3.1) involves using variables (U_l, M_l) and formulating a complete-data likelihood. An expectation-maximization algorithm is then employed, with each iteration of the M-step using closed form updating equations for π_w. The parameter v is updated using the quasi-Newton Raphson method known as Broyden-Fletcher-

Goldfarb-Shanno (Broyden, 1970), based on the expected complete-data log-likelihood given the observed data. Multiple random initialization of v are utilized. Additionally, π_w is initialized with a uniform distribution, and $\epsilon = 10^{-3}$.

cc	kk	entropy	LL	AIC	BIC	q	alloc
4	1	1.26	-62.56	-137.12	-144.1883	0.18,0.46,0.25,0.11	1,9,11,3
3	1	0.88	-68.25	-144.50	-149.2122	0.63,0.25,0.11	10,11,3
3	2	0.95	-68.25	-144.50	-149.2122	0.52,0.37,0.11	10,11,3
3	1	1.04	-77.74	-163.48	-168.1922	0.29,0.49,0.23	3,10,11
3	2	0.75	-77.74	-163.48	-168.1922	0.06,0.23,0.71	3,11,10
2	1	0.65	-81.38	-168.76	-172.2942	0.64,0.36	9,11
2	1	0.56	-86.32	-176.64	-178.9961	0.75,0.25	10,14

Table 3.2: SMASH output

Table 3.2 demonstrates the output from SMASH clonality analysis. Here the "cc" column refers to the number of subclones and the "kk" column refers to the topology of the tree. For example cc = 4, kk = 1 corresponds to the linear tree with 4 subclones, while cc = 3, kk = 2 corresponds to a nonlinear tree with number of clones equal to 3. The strings in the "alloc" column, for example $1|1; 2|9; 3|11; 4|3$, indicates that 1 mutation emerged in the first subclone (founding clone), 9 mutations in the second subclone, 11 mutations in the third subclone, and 3 mutations in the last subclone. Also, $q = (0.18, 0.46, 0.25, 0.11)$ gives the estimates for subclone proportions (v_s). "entropy" column is the entropy that we refer in this thesis as proportion entropy (PE) :

$$PE = -\sum_{s=1}^{S} q_s \log(q_s), \tag{3.2}$$

where q_s are elements of vector q.

3.2.1 Functional clonality

In our investigation of tumor evolution, we recognize that not all gene mutations contribute significantly to the progression of liver tumors (Dressler et al., 2022). Therefore, we employ a subset of gene mutations within the SMASH algorithm that are likely to play a functional role in liver tumor development. Specifically, we select 1006 functional genes, out of which 386 genes are known to be over-expressed in normal liver tissue. The remaining genes are previously identified as potential drivers of liver cancer. We adopt this focused approach in all three cohorts described in Section 1.7.

To better understand the influence of functional clonality on clinical outcomes, we conducted a comparative analysis of survival distributions. We examined the linear and nonlinear phylogenies resulting from two distinct analyses: clonality analysis focusing on functional gene mutations and clonality analysis encompassing all gene mutations. By examining the differences in survival outcomes, we can evaluate the efficacy of functional clonality in revealing distinct patterns associated with clinical prognosis.

Figure 3.1 displays the comparison results with 95% confidence intervals for the hazard ratio of linear versus nonlinear trees in the TCGA-LIHC and TIGER-LC cohorts. The hazard ratio estimates were obtained using the coxph function from the R package, fitting the Cox proportional hazards model. The coxph function employs partial likelihood estimation to account for censored observations in survival data.

The hazard ratio represents the relative risk of an event, such as death, between two groups. In this analysis, the hazard ratio was calculated to assess the impact of tree structure (linear versus nonlinear) on survival outcomes. The forest plot provides a visual representation of the hazard

ratio estimates, allowing for a comparison of the effect of tree structure on patient subgroups. These results suggest that targeting functional mutations is a more effective strategy than targeting all mutations in identifying patient subgroups with different survival outcomes. The hazard ratio estimates obtained through the Cox model analysis provide evidence supporting this conclusion.

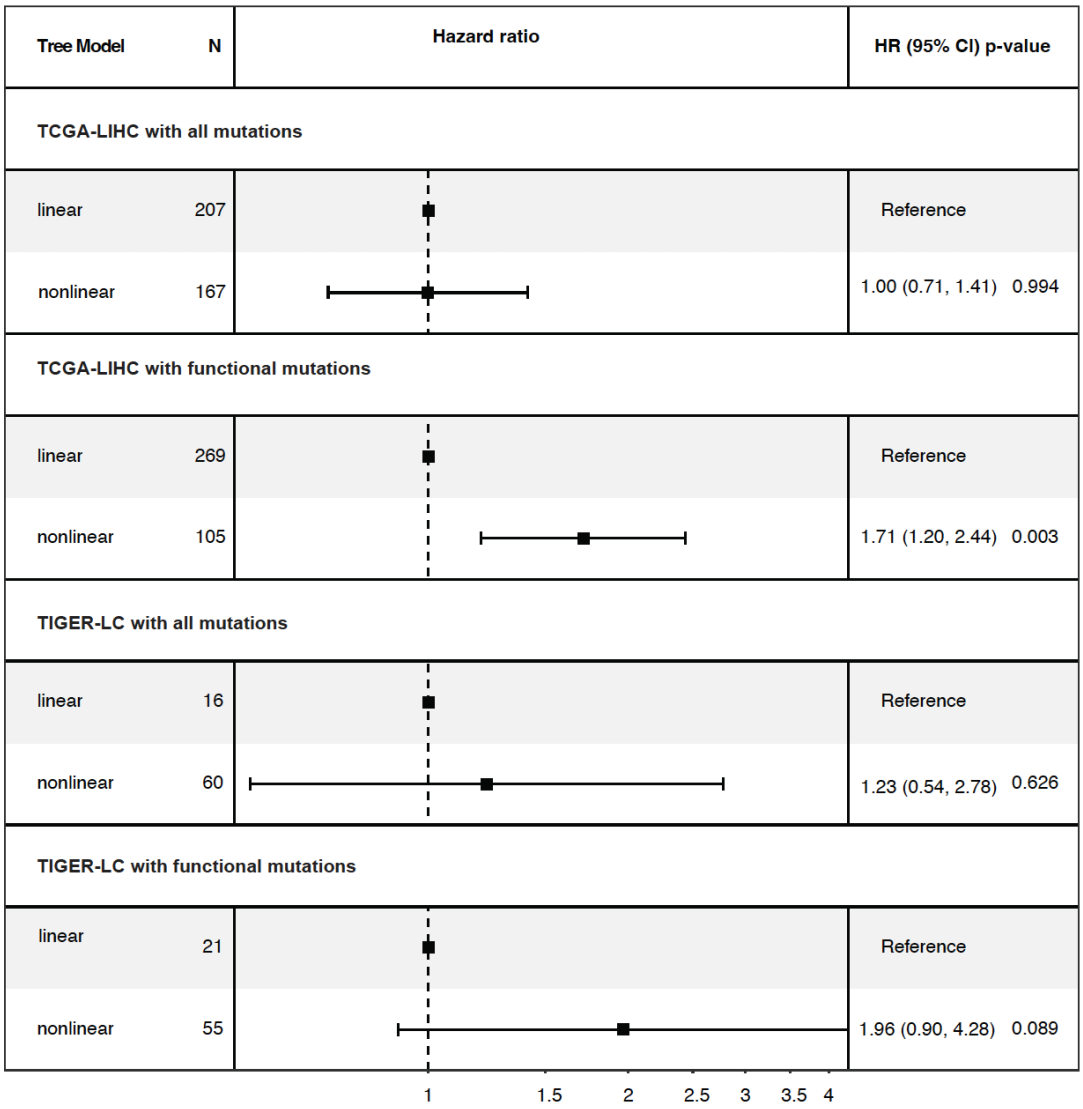


Figure 3.1: Forest Plots. The hazard ratios with their corresponding 95% confidence intervals for the comparison between functional mutations and all mutations. In this analysis, our primary interest lies in identifying hazard ratios that deviate significantly from 1.0

3.2.2 Linear versus nonlinear tumor phylogenies

The first and most well-known model of tumor evolution is the linear evolution model based on the ideas of Nowell (Nowell, 1976), where tumors accumulate clonal mutations with highly dominant selective properties, outcompeting all previous clones. For a long time, tumor evolution was only believed to be a linear accumulation of clonal mutations. However, observations in several studies (Dexter et al., 1978; Heppner, 1984) showed the possibility of nonlinear growth for tumors with several molecularly distinct subclones. The SMASH algorithm identifies linear or nonlinear trees, where linear trees have a dominant clone that progresses sequentially, and nonlinear trees have multiple major clones that evolve in parallel during tumor progression (Davis et al., 2017). This work further classifies nonlinear trees as either deep or shallow branching. Several recent papers, Davis et al. (2017); Vendramin et al. (2021); Zhu et al. (2021), review different tumor evolution models with their distinct biological features including but not limited to linear evolution, branching evolution, neutral evolution, and macroevolution.

In Figure 3.2, we display a visual representation of linear (a) and nonlinear (b,c) tumor trees from TCGA-LIHC cohorts. For example, for the linear tree in (a), founding clone accounts for 61% of tumor cells and harbors two mutations. Subsequently, mutations accumulate sequentially, resulting in the formation of subclones 1, 2, and 3. The numbers adjacent to the branches indicate the number of unique mutations in the corresponding subclone, and the branch lengths are proportional to this number. In the nonlinear tree (b), on the other hand, subclone 2 and 3 for tree have different sets of mutations since they evolved independently in parallel.

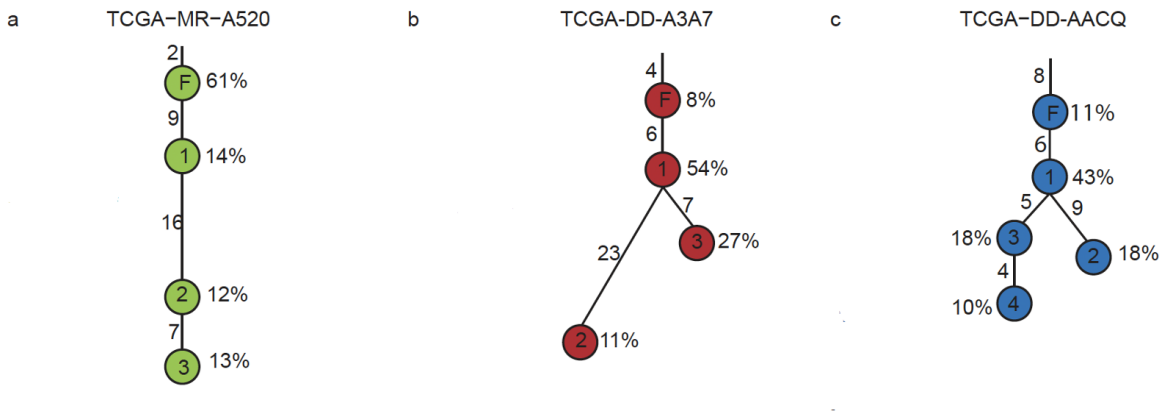


Figure 3.2: Linear versus Nonlinear Trees

3.3 Clustering analysis of nonlinear tumor phylogenies

In the literature, two common types of nonlinear evolution models have been proposed based on the presence or absence of selection advantages among subclones (Zhu et al., 2021). Those results may be compatible with our resulting nonlinear tumor trees, especially two popular ones, branching and neutral. Therefore, we further investigate nonlinear trees to see if there are two subgroups of nonlinear trees to which we can assign biologically meaningful labels. To distinguish distinct tree clusters within the nonlinear class, we develop a clustering algorithm that initially creates several diversity variables (entropies) of each tree using subclone proportions and proportions of mutations by subclone and then uses these created variables to cluster and label trees based on how balanced they are, i.e., based on how similar the sizes and numbers of mutations in the subclones are. It classifies nonlinear trees as shallow branching or deep branching, depending on their structure. Branching trees with similarly sized branches and subclones are classified as deep branching, and those without this feature are designated shallow branching.

3.3.1 Feature creation

To maximize the information extracted from the outputs of the clustering process, we aim to capture important topological features of the tumor phylogeny trees. To achieve this, we define and calculate several features utilizing subclone proportions and mutation probabilities. We define these features as follows. Let S be the total number of subclones in a clonal tree and q_s be the estimated proportion of the s^{th} subclone. Also, denote by m_s as the relative frequency of the unique mutations in the s^{th} subclone. Then, the proportion entropy (PE) is same as in Equation (3.2) and mutation entropy (ME) is defined by:

$$ME = -\sum_{s=1}^{S} m_s \log(m_s). \tag{3.3}$$

In addition to these entropies, we create 7 more features, which are given in Equations (3.4)-(3.10) by renormalizing the vectors after taking component-wise ratio and product of two entropies, as well as taking squares and cubes of the entries of two entropies. Let $Q = (q_1, q_2, \ldots, q_S)$ be the subclone proportion vector and $M = (m_1, m_2, \ldots, m_S)$ be the mutation frequency vector with q and m being the means of Q and M.

$$cor = \frac{\sum_{i=1}^{S}(q_i - q)(m_i - m)}{\sqrt{\sum_{i=1}^{S}(q_i - q)^2(m_i - m)^2}}, \tag{3.4}$$

$$E_{raito} = -\sum_{i=1}^{s} \frac{q_i/m_i}{\sum_{i=1}^{s} q_i/m_i} \log\left(\frac{q_i/m_i}{\sum_{i=1}^{s} q_i/m_i}\right), \tag{3.5}$$

$$E_{product} = -\sum_{i=1}^{s} \frac{q_i m_i}{\sum_{i=1}^{s} q_i m_i} \log\left(\frac{q_i m_i}{\sum_{i=1}^{s} q_i m_i}\right), \tag{3.6}$$

$$ME_{square} = -\sum_{i=1}^{s} \frac{m_i^2}{\sum_{i=1}^{s} m_i^2} \log\left(\frac{m_i^2}{\sum_{i=1}^{s} m_i^2}\right), \tag{3.7}$$

$$ME_{cube} = -\sum_{i=1}^{s} \frac{m_i^3}{\sum_{i=1}^{s} m^3} \log\left(\frac{m_i^3}{\sum_{i=1}^{s} m_i^3}\right), \tag{3.8}$$

$$PE_{square} = -\sum_{i=1}^{s} \frac{q_i^2}{\sum_{i=1}^{s} q_i^2} \log\left(\frac{q_i^2}{\sum_{i=1}^{s} q_i^2}\right), \tag{3.9}$$

$$PE_{cube} = -\sum_{i=1}^{s} \frac{q_i^3}{\sum_{i=1}^{s} q_i^3} \log\left(\frac{q_i^3}{\sum_{i=1}^{s} q_i^3}\right). \tag{3.10}$$

As a final step in feature creation, we normalize each of the entropy features by $\log(S)$, which is the largest possible entropy with S number of subclones, in each tumor tree to account for the number of subclone differences (normalized entropy). This choice to normalize by S effectively prevents the number of subclones from serving as a predictive feature for clustering, but we also conducted a similar analysis without normalizing by S and found that the unnormalized method produces similar results.

3.3.2 Clustering algorithm

In unsupervised learning, such as clustering, the data consists of a uniform set of features without any class labels. Initially, we explored popular algorithms like k-means and hierarchical clustering, as discussed in Section 1.9, for clustering nonlinear trees using the aforementioned features. However, these algorithms did not yield satisfactory and consistent results across all cohorts, particularly under re-sampling scenarios. To overcome this limitation and identify an optimal dissimilarity measure, we employed various methods, as described in Section 1.8, in particular one based on random forests.

Random forests can be used for clustering (Breiman, 2001; Breiman and Cutler, 2023) by

treating the original data as class 1 and creating a synthetic second class of the same size that is labeled as class 2. Consequently, the augmented dataset with two class labels can be used in the random forest algorithm. The primary goal of labeling the original and augmented data is to generate a similarity measure that can be used to assess the similarity between pairs of original data points. By randomly sampling from the univariate distributions of the original features, the augmented dataset effectively eliminates the cross-coordinate dependency structure present in the original data. This process ensures that the synthetic second class is not influenced by the inherent relationships and dependencies among the features in the original data.

Once the augmented dataset is constructed with two class labels, it can be used as input to the random forest algorithm. Random forests operate by building an ensemble of decision trees, where each tree is trained on a random subset of the data. In the clustering context, the random forest algorithm aims to learn the patterns and relationships within the data that distinguish class 1 (original data) from class 2 (synthetic data). The resulting random forest model captures the underlying structure and patterns in the original data, providing a similarity measure that can be used to assess the similarity between pairs of original data points. This similarity measure is not influenced by the synthetic second class but rather focuses on the intrinsic patterns within the original data.

In this random forest clustering algorithm, we calculate the proximity of a pair of observations (similarity measure) in the following way: grow many (10000) random forest trees independently, count the number of times those observations ended up in the same terminal node, and finally normalize this quantity by the number of total trees. Finally, we use this new similarity matrix in the k-means algorithm with the following specified settings: centers=2, nstart=20, iter.max=10. The k-means algorithm is implemented with randomly chosen centers and is run many times to

get consensus results. Thus, the algorithm extracts 2-class data augmentation 20 times, run the algorithm 10 times for each data, and choose the optimal clustering as the one with the lowest total sums of squares (distances in all the clusters of the points from the centroids). To assign labels to resulting clusters, we use the fact that the deep branching trees, which resemble the neutral trees in literature, are the extreme cases, so we would expect them to have a larger entropy of the terminal node size distribution than shallow branching trees. This labeling strategy is used to align the clusters. For this reason, we calculate the average proportion and mutation entropy of each cluster and find the maximum of these two quantities, finally, the cluster with the higher maximum is assigned as deep branching. We denote mean mutation entropy by ME, proportion entropy by PE, in each case with subscript indicating cluster, and then define

$$m_1 = \max(ME_{cluster1}, PE_{cluster1}), \tag{3.11}$$

$$m_2 = \max(ME_{cluster2}, PE_{cluster2}). \tag{3.12}$$

After assigning the clusters based on the maximum value of $m = \max(m_1, m_2)$, we label the cluster with the larger maximum as "deep branching", while the other cluster is labeled as "shallow branching". Using these cluster labels and the features created in the previous section, we performed Principal Component Analysis (PCA) on the data. Before conducting PCA, the data was standardized by subtracting the mean and dividing by the standard deviation of each variable. This standardization ensures that each variable contributes equally to the analysis and prevents any single variable from dominating the results.

The resulting PCA plots, shown in Figure 3.3, provide a visual representation of the clusters

obtained from the analysis. These plots allow us to observe the separation between the "deep branching" and "shallow branching" clusters and further validate the effectiveness of the clustering approach.

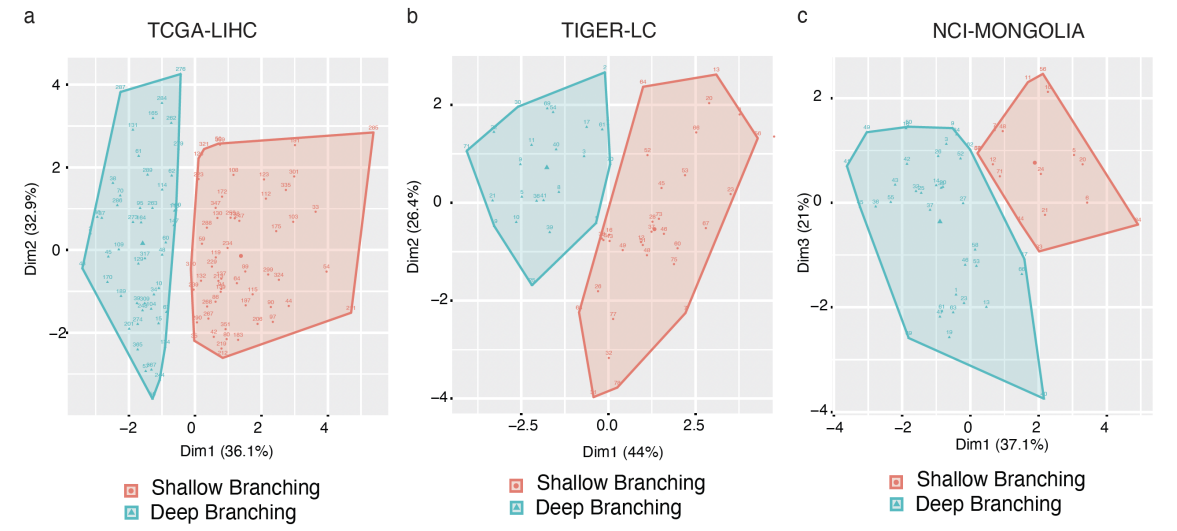


Figure 3.3: PCA plots for tumor tree clusters

3.3.3 Random forest proximities as a measure of similarity

As discussed in Section 1.9.1, machine learning techniques can be trained to learn dissimilarity measures directly from the data (Xing et al., 2002). These approaches, often referred to as metric learning or distance metric learning, aim to optimize a distance function that captures the dissimilarity between samples based on the clustering objectives. In our clustering analysis, we adopt a similar approach by utilizing random forest to establish a measure of similarity between features, even though random forest is primarily used for supervised learning. Figure 3.4 demonstrates the relationship between calculated proximities and Euclidean distance of observation pairs for each of our cohorts. The blue curves in the generated plots were computed using the geom_smooth function with the method set to ”loess” (locally estimated scatterplot smoothing).

This method fits a smooth curve to the scatterplot data, capturing the underlying trend in the relationship between the distances and proximities.

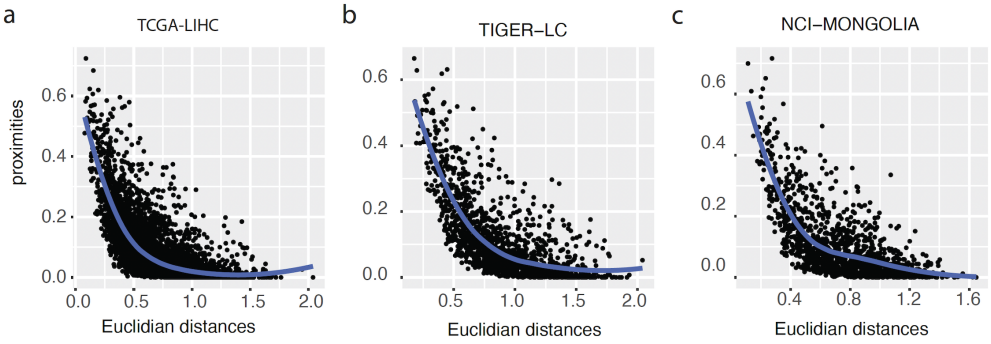


Figure 3.4: Proximity versus Euclidean distance. Each dot represents the pair of observation from non-linear trees.

The concept of proximity differs from the traditional Euclidean distance commonly used in clustering algorithms. Rather than relying solely on the geometric distance between data points, the random forest proximities encapsulate additional information captured through the decision tree structure. This information accounts for complex relationships and interactions among the features, providing for this specific problem at hand a more comprehensive dissimilarity measure.

3.3.4 Stability analysis of clustering algorithm

To ensure the stability of our clustering analysis, we conduct a stability analysis to demonstrate the stability of our shallow and deep branching tree clusters under bootstrap re-sampling.

To assess the stability of our clustering analysis, we perform a bootstrap analysis. This analysis aims to demonstrate the robustness and consistency of our identified shallow and deep branching tree clusters across multiple iterations of resampling. We employ non-parametric bootstrapping, where the rows are re-sampled while retaining the fully preprocessed feature observations. The random forest clustering approach discussed above is then applied to each

bootstrapped data set, generating proximity values. These proximities are subsequently utilized in the k-means clustering algorithm. We compare the clustering results obtained from the bootstrap data with those from the original data by calculating the affinity score (Mainali et al., 2022) based on the 2x2 frequency table. This process is repeated for a total of $B = 1000$ bootstrap replications. The Algorithm 1 outlines the steps involved in the stability analysis. The estimated average log odds-ratios for each bootstrap affinity score is found to be 5.45 for TCGA-LHC, 6.02 for TIGER-LC, and 6.64 for NCI-MONGOLIA, indicating high agreement between cluster results. The histogram of estimated affinity scores, along with their corresponding p-values, is provided in Figure 3.5.

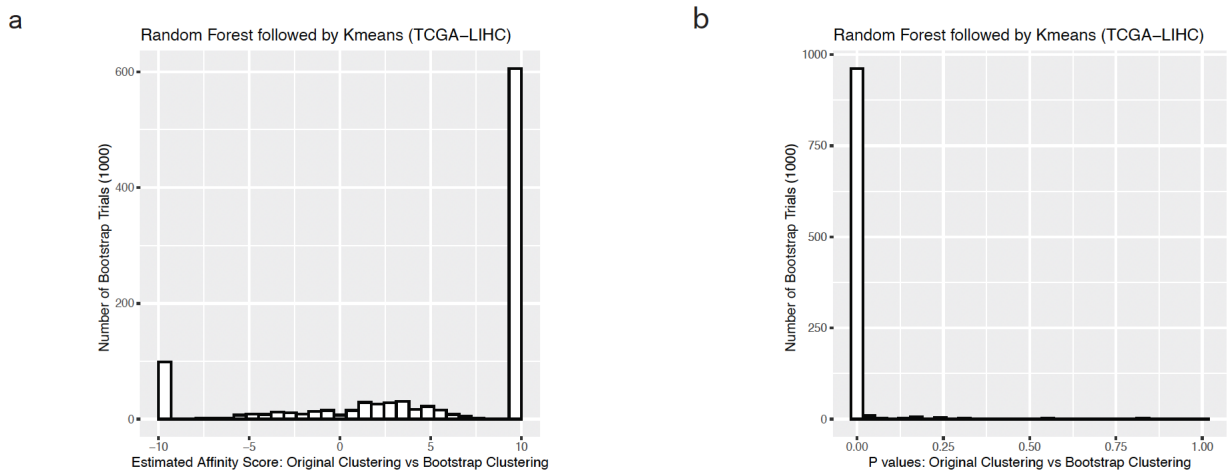


Figure 3.5: Affinity Score Estimates (a) and P-values (b)

This stability analysis provides evidence that our shallow and deep branching tree clusters are stable under bootstrap resampling, supporting the reliability of our clustering results.

for *b = 1 to 1000* **do**

Bootstrap resampling: Draw a bootstrap sample D^* of the same size N as the original data by randomly selecting rows from the data with replacement;

Synthetic data creation: Create another synthetic data D^{**} from D^* by sampling from univariate distributions of features in D^*. Each element of a row of D^{**} comes independently from the original features;

Random forest training: Run a random forest on a problem with two labels that represent D^* and D^{**} respectively;

Proximity calculation: Calculate the proximities of a pair of observations in D^*. To do this, grow 1000 random forest trees on the augmented bootstrap-replicated data (D^*, D^{**}) and calculate the overall (normalized) proximity between nodes i and j of D^*. The proximity is proportional to the count of these trees in which i and j fell in the same cluster;

Cluster alignment: Feed the proximity matrix obtained in the previous step to the k-means algorithm. Determine the label of each row as either shallow or deep branching using the values of m_1 and m_2;

Clustering comparison: For each row of D^*, compare the clustering results for the original data and the bootstrap sample. Calculate the affinity score (similarity measure) of the 2x2 frequency table;

end

Algorithm 1: Stability Analysis of the Clustering Algorithm Defining "deep branching" and "shallow branching" Tumor Trees

3.4 Implications of clonality and clustering analysis results

The clonality and clustering analysis of this study reveal the existence of three distinct phylogenetic tree groups within HCC cohorts: linear, shallow branching, and deep branching. Shallow branching and deep branching trees differ in that deep branching trees show a relatively high degree of balance with respect to edge-lengths (numbers of new mutations) and prevalence (read-counts). For the patient trees in Figure 3.2 (b, c), the proportion entropy (PE) and the mutation entropy (ME) were both 1.14 for shallow branching tree (red), whereas for the deep branching (blue) tree PE is 1.45 and ME is 1.57.

The results of our analysis suggest that the majority of the TCGA-LIHC cohort consists of linear trees (270 linear, 56 shallow branching, 49 deep branching). On the other hand, the TIGER-LC cohort displays a higher proportion of shallow and deep branching trees (22 linear, 32 shallow branching, 24 deep branching), while the NCI-MONGOLIA cohort exhibits a similar trend with 18 linear, 18 shallow branching, and 35 deep branching trees.

These preliminary findings provide some understanding into the clonal evolution patterns within HCC. However, it is important to approach these cluster groupings with caution and acknowledge the need for further validation. To strengthen the validity of our findings, additional analyses and validation checks will be conducted in the subsequent sections and Chapter 4. These subsequent analyses, including survival analyses, will help further evaluate the biological relevance and implications of the identified clusters in terms of tumor progression, driver gene profiles, and the impact of the tumor microenvironment on cancer development and treatment response.

3.4.1 Survival outcomes of tumor evolution phylogenies

To investigate the impact of tumor evolution on survival, we examine the Kaplan Meier (Kaplan and Meier, 1958) survival plots for groups of subjects defined by phylogenetic clusters. We find that overall survival for trees that are linear have statistically better prognosis compared to the nonlinear ones although this tendency is less clear in the smaller NCI-MONGOLIA cohort than in he other cohorts.

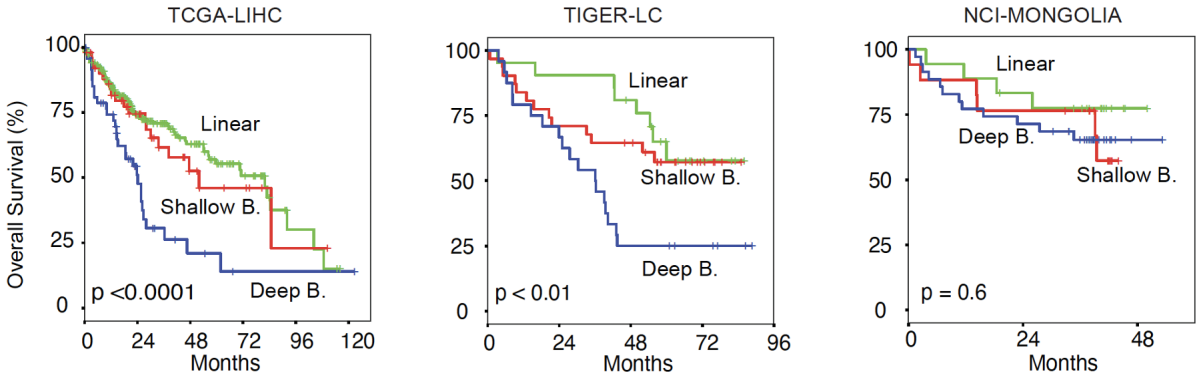


Figure 3.6: Kaplan Meier Survival Curves: Linear, Shallow Branching and Deep Branching

In addition to the forest plots shown in Figure 3.1, we provide an additional figure that visually represents the Kaplan-Meier curves for both linear and nonlinear trees in the TCGA-LIHC and TIGER-LC cohorts. This figure presents the results of two separate analyses, highlighting the differences in survival outcomes between the two tree types. Functional mutations are specifically associated with liver or liver cancer-related genes. By focusing on these mutations, we are targeting alterations that are likely to have a more direct impact on the biological processes underlying tumor progression and patient prognosis. The significant separation in survival curves observed between linear and nonlinear trees when considering only functional mutations (Figure 3.7, A, C) suggests that the specific tree structure becomes more influential in identifying distinct

prognostic subgroups when focusing on functionally relevant mutations.

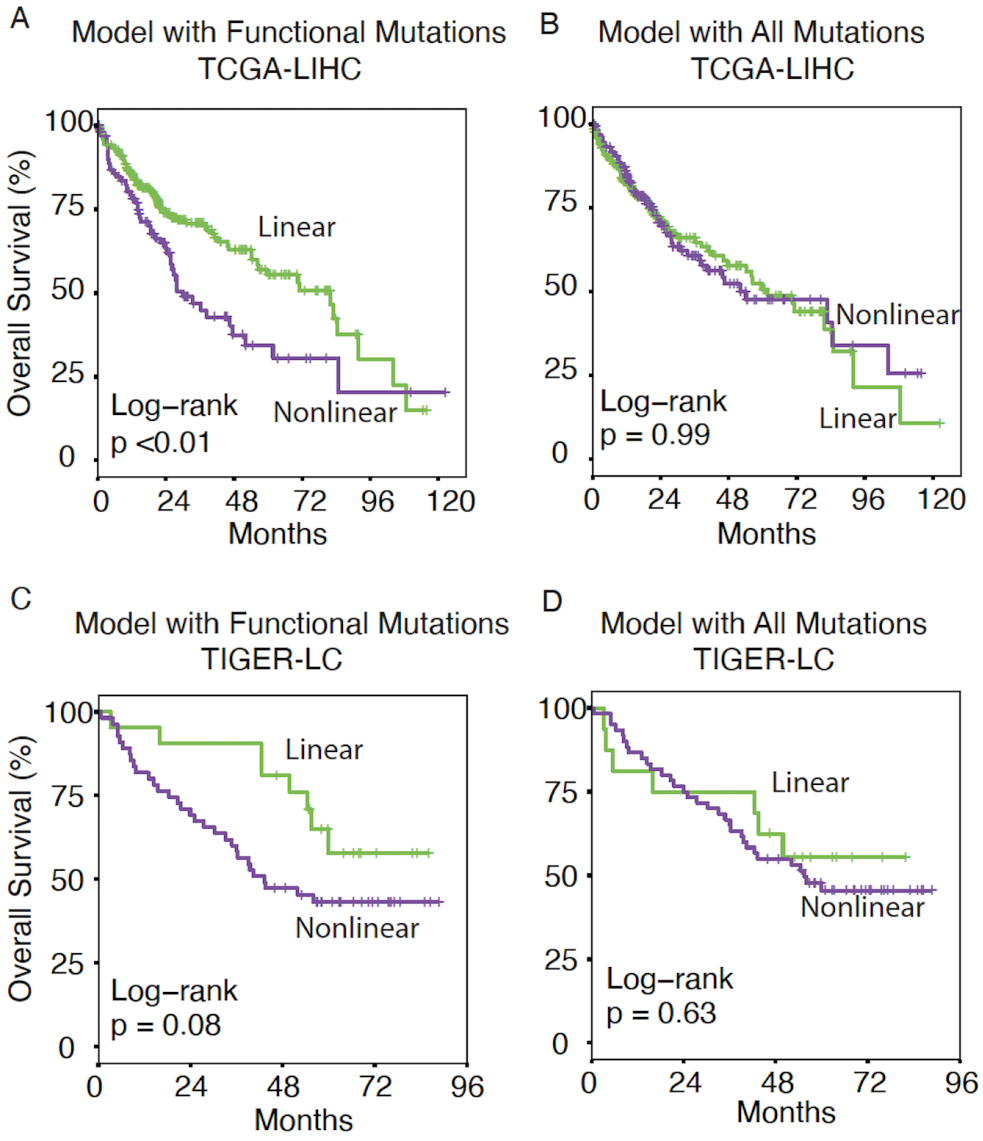


Figure 3.7: Kaplan Meier Survival Curves of Linear versus nonlinear. A,C : Only functional mutations are utilized. B,D: All mutations are utilized

3.4.2 Potential drivers of tumor evolution phylogenies

One possible explanation for the diverse paths of tumor evolution observed in HCC tumors could be the presence of distinct driver mutations for each type of tumor evolution phylogeny. To explore this hypothesis, we compare the mutation profiles of linear, shallow branching, and deep

branching phylogenies. We define a gene as a potential driver if it is mutated in the founding clone.

Figure 3.8 presents potential driver profiles of linear, shallow branching, and deep branching trees in TCGA-LIHC (top row), TIGER-LC (second row), and NCI-MONGOLIA (third row) cohorts. In all cohorts, deep branching trees exhibited the highest rate of driver mutations in TP53, whereas in linear trees, TP53 was not the most mutated gene. Notably, for nonlinear phylogenies, TP53 and CTNNB1 were the two most frequently mutated genes, whereas for linear phylogenies, they did not even rank within the top 5 in the TIGER-LC and NCI-MONGOLIA cohorts. In the TCGA-LIHC cohort, these gene mutations remained dominant in nonlinear trees, but in linear trees, MUC6 was the most frequently mutated gene across all patients. While this trend was observed, no significant correlation was found between TP53 mutation and phylogenetic model evolution. Conversely, the mutation frequencies of the GTF2IRD2B gene showed an opposite trend, suggesting that it may be a driver of linear evolution, particularly in the NCI-MONGOLIA cohort. To validate this, we performed the Freeman-Halton extension (Freeman and Halton, 1951) of the Fisher's exact test for 3x3 contingency tables to check the association between mutation status of GTF2IRD2B (no mutation, driver, not driver) and tree evolution model (linear, shallow branching, deep branching). In all cohorts, the exact test p-values (< 0.01) indicated that the GTF2IRD2B gene status was associated with the tree evolution phylogeny, suggesting its potential role as a driver for linear phylogenies.

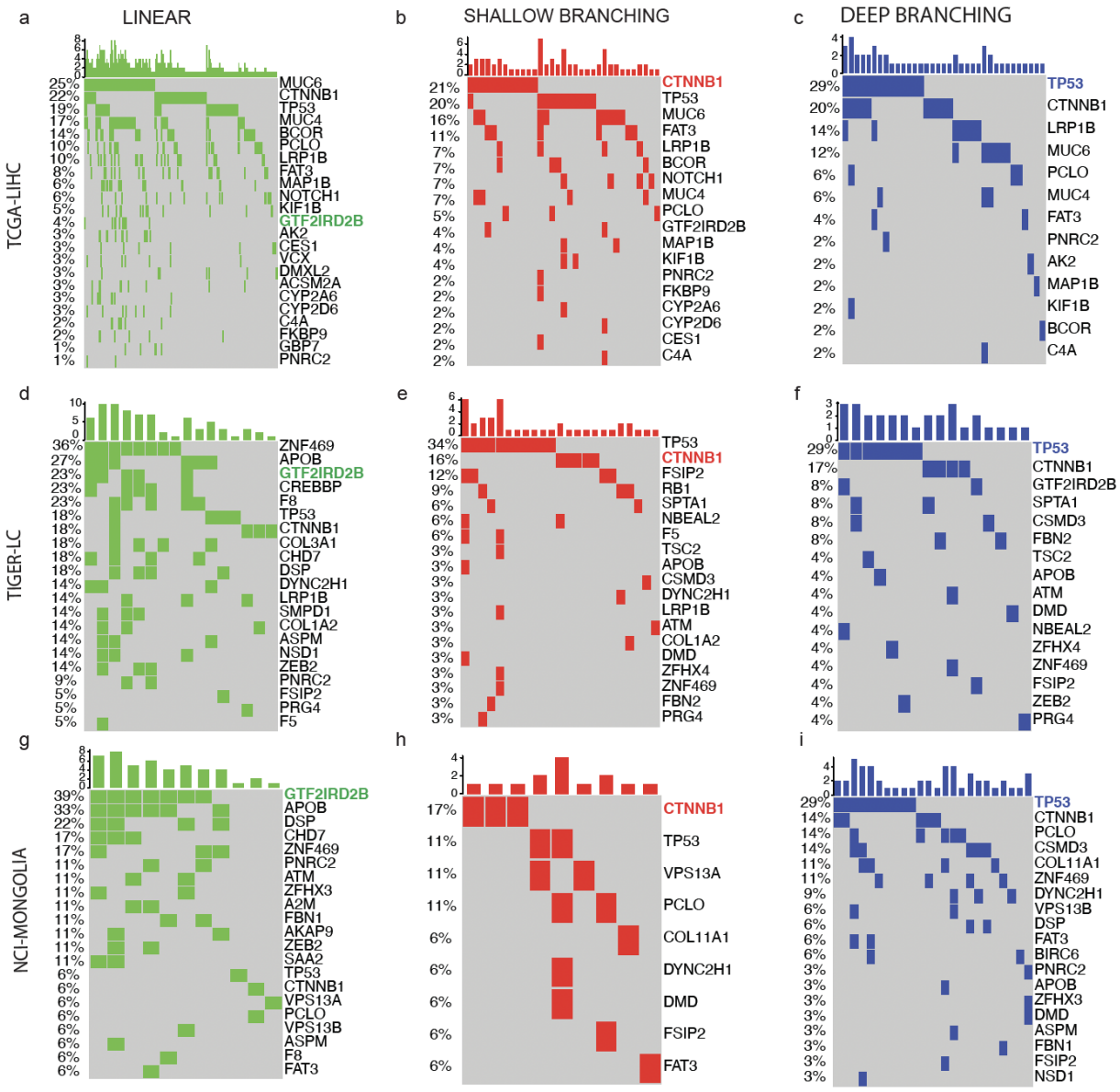


Figure 3.8: Driver Gene Profiles of Tumor Phylogenies

3.4.3 Immune cell micro-environment of tumor evolution phylogenies

The interplay between tumor cells and their microenvironment is critical in tumor progression and treatment response. It has been demonstrated that the clonal architecture of tumors can shape their microenvironment (Zhang et al., 2021). To investigate this further, we utilized RNA-sequencing data from the same tumor samples and performed a transcriptome analysis using

CIBERSORTx, a deconvolution (computational technique used to estimate the relative proportions of different cell types within a complex tissue sample) tool (Steen et al., 2020).

3.4.3.1 Immune Cell Decomposition: CIBERSORTx

CIBERSORTx is used for characterizing cell composition in complex tissues based on gene expression data. It is designed for analyzing bulk RNA sequencing data as well as single cell RNA sequencing data to estimate the abundance of different immune cell types within a heterogeneous sample. We use the count matrix data for tumor samples of a cohort that is shown as in 1.5 as an input to CIBERSORTx. The usage of CIBERSORTx involves several steps.

1. Input Data: To prepare the input data for CIBERSORTx, bulk RNA-seq expression data should be in a specific format. This data should include the gene expression profiles of your samples and a reference gene expression signature matrix representing the different cell types of interest. In this study, we utilize the LM22 reference signature matrix (Newman et al., 2015), which consists of 547 carefully selected genes. This matrix has been specifically designed to accurately differentiate 22 distinct human immune cell populations. The development and extensive validation of the LM22 matrix were performed using gene expression micro-array data. However, it is also applicable for RNA-Seq data analysis, allowing for the generation of hypotheses and exploration of gene expression patterns. The 22 cell populations encompass various T cell types, including both naive and memory B cells, plasma cells, natural killer (NK) cells, and different subsets of myeloid cells. For the specific names of these 22 cell types, refer to Figure 3.9, where they are provided as x-labels in the plot.

2. Run CIBERSORTx: Upload input data to the CIBERSORTx web portal or use the command-

line version of the tool. The tool will perform deconvolution analysis using a machine learning algorithm to estimate the proportions of different cell types within the samples.

3. Analysis Parameters: Specify the necessary parameters for your analysis, such as the reference signature matrix, the number of permutations for statistical testing, and any other relevant options.

4. Deconvolution Results: Once the analysis is complete, CIBERSORTx provides the estimated abundance of each cell type in the samples. The output consists of rows representing the samples and columns representing the 22 cell types, with each value representing the estimated frequency of the respective cell type. The frequencies of the cell types within each sample sum up to 1, indicating the proportional composition of the cells. These results can be visualized and subjected to further analysis to gain a deeper understanding of the tissue or sample composition.

Further investigation of myeloid and lymphoid cells separately is motivated by the fact that these two cell populations have distinct functions and characteristics within the immune system. Myeloid cells encompass a variety of immune cell types, including macrophages, dendritic cells, and granulocytes. They are involved in innate immunity and contribute to tissue homeostasis, inflammation, and immune responses. In the tumor microenvironment, myeloid cells can have both pro-tumorigenic and anti-tumorigenic effects, depending on their polarization state and functional properties. On the other hand, lymphoid cells are primarily responsible for adaptive immune responses and include various types of T cells, B cells, and natural killer (NK) cells. These cells play a crucial role in recognizing and targeting cancer cells, orchestrating immune responses, and generating immunological memory.

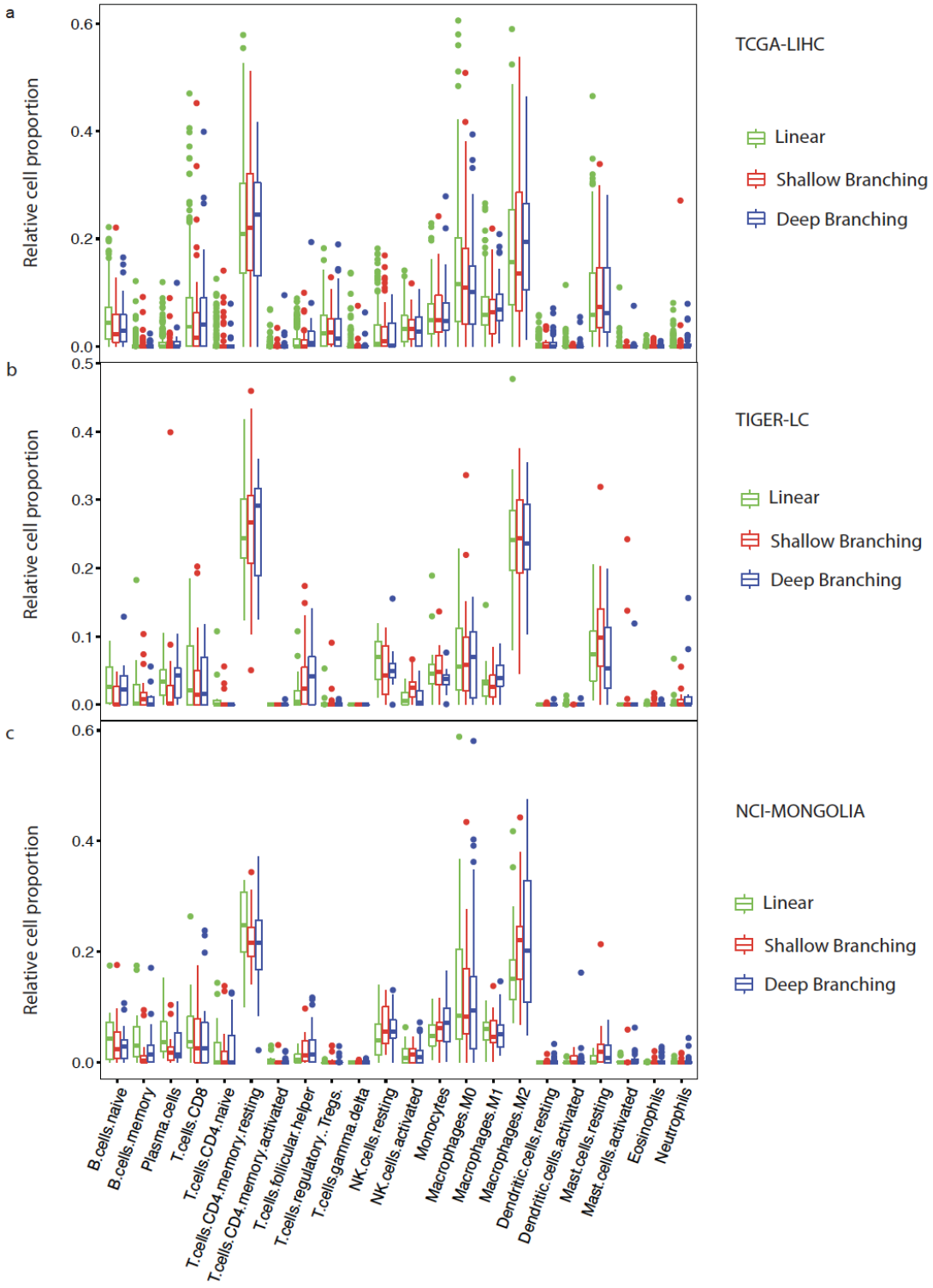


Figure 3.9: Tumor Micro-environment of Tumor Phylogenies for TCGA-LIHC (a), TIGER-LC (b), and NCI-MONGOLIA cohorts. Each of the 22 cell types listed on the x-axis is associated with three box plots representing different measurements of individuals with linear, shallow branching or deep branching trees. The y-axis represents the relative proportion of these cell types within the tumor micro-environment. The box plots provide information into the distribution and variability of the relative proportions of each cell type across different phylogenies.

The investigation of myeloid and lymphoid cell frequencies within the tumor microenvironment using a tumor evolution model yielded intriguing results. Specifically, the analysis revealed that in all three cohorts studied, tumor evolution models characterized by shallow branching trees exhibited higher levels of myeloid cell frequencies and lower levels of lymphoid cell frequencies compared to linear models (Figure 3.10). This unexpected finding suggests that the chosen tumor evolution model may influence the relative abundance and distribution of immune cell types within the tumor microenvironment.

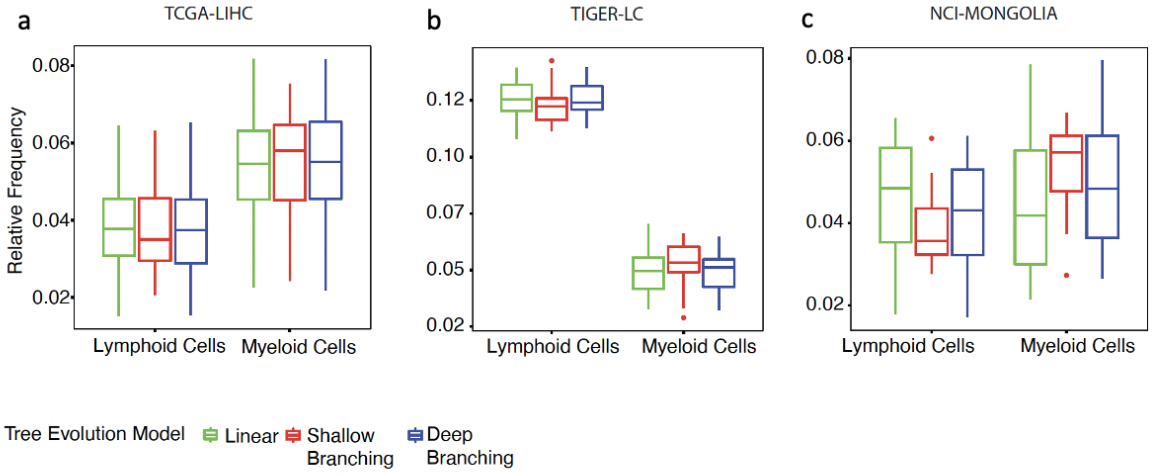


Figure 3.10: Myeloid and Lymphoid Cell Frequencies for TCGA-LIHC (a), TIGER-LC (b), and NCI-MONGOLIA cohorts. Each box plot represents the distribution of cell frequencies for either myeloid or lymphoid cells within three distinct tree phylogenies (linear, shallow branching, deep branching). The colors assigned to the box plots differentiate the tree phylogenies within each cohort.

After observing elevated levels of lymphoid cells in linear tree tumors, we focuse our investigation on B cells due to their crucial role in the immune response against cancer. Our results, depicted in Figure 3.11, demonstrate that the total B cell frequencies are significantly higher in linear trees compared to nonlinear trees for all cohorts (Kruskal-Wallis's test p-values ≤ 0.05).

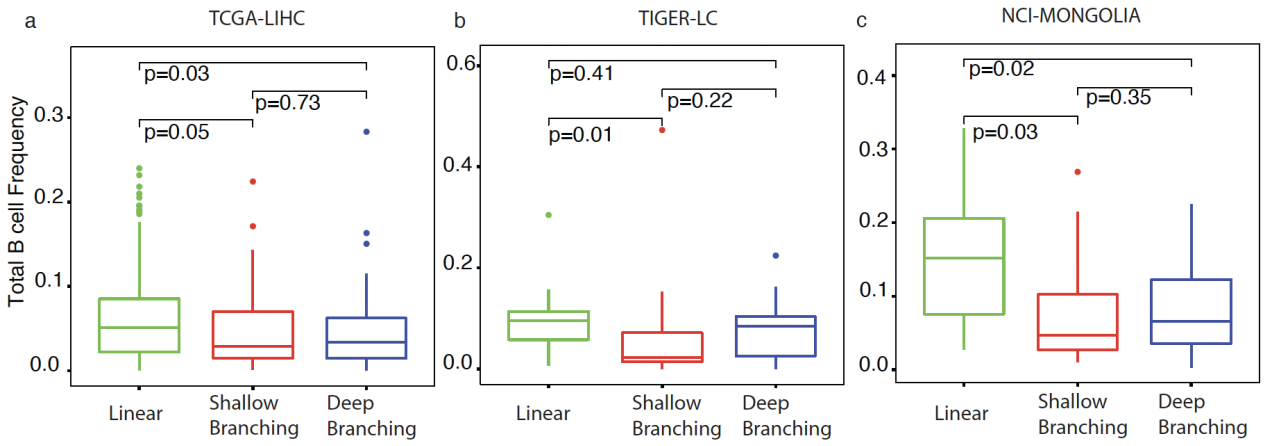


Figure 3.11: B Cell Frequencies for TCGA-LIHC (a), TIGER-LC (b), and NCI-MONGOLIA cohorts. Each box plot represents the distribution of Total B Cell frequencies for three distinct tree phylogenies, linear, shallow branching, and deep branching.

3.5 Discussion

Intra-tumor heterogeneity (ITH) is the phenomenon of clonal variability within a patient's tumor, which arises as a result of stochastic (mutation, drift) and deterministic (selection) processes in the evolution of cancer. In this study, we represented ITH by reconstructing the clonal trees and calculating various features using mutation entropy as well as the proportion entropy that previous studies used. By combining these representations of ITH with phylogenetic tree construction, we successfully compared the biological and clinical paths of tumor evolution. A typical tumor evolution analysis combines single nucleotide variants (SNVs) and copy number alterations (CNAs), but not all SNVs are relevant to cancer progression. Moreover, many phylogenetic reconstruction algorithms are unreliable when the number of mutations is high. To deal with redundancy in mutations, we focus on liver and liver cancer specific genes, successfully showing that functional mutations improve clonality analysis.

Some somatic alterations in specific genes, known as "driver genes," contribute to tumorigenesis

by granting selective advantages to certain tumor cells (Stratton et al., 2009). This study aims to find drivers of each tumor phylogeny by examining the founding clone mutations in each phylogeny. The GTF2IRD2B gene is a potential driver for a linear phylogeny in hepatocellular carcinoma (HCC), while TP53 and CTNNB1 are candidates for driving more aggressive branching phylogenies.

The tumor microenvironment (TME) plays a role together with tumor cells in tumor progression and response to treatment. Successful establishment of tumor clonality requires a comprehensive understanding of the development of somatic alterations in tumor cells and the formation of a conducive TME that facilitates the survival and growth of these altered tumor cells (Ma et al., 2022). Thus, the interaction between tumor cells and their microenvironment validates a successful clonality construction. In this study, we aimed to provide evidence for such an interaction. We first focused on B cells in our study because they have been shown to play a crucial role in the immune response against cancer. B cells can produce antibodies that target tumor antigens, leading to their destruction by other immune cells. Additionally, B cells are involved in antigen presentation and immune regulation, which can impact the overall anti-tumor immune response. Our results showed the frequency of B cells is significantly higher in linear trees, which are associated with less aggressive tumors, compared to nonlinear trees. This suggests that B cells may play a role in suppressing tumor progression in less aggressive tumors, while they are less effective in more aggressive tumors. These findings are consistent with previous studies that have shown a positive association between the presence of B cells in tumors and improved patient outcomes in HCC (Zhang et al., 2019). Furthermore, our observation that nonlinear trees had the highest myeloid cell frequencies and the lowest lymphoid cell frequencies compared to linear in all three cohorts is also consistent with studies demonstrating a negative correlation between

myeloid cells and tumor prognosis (Engblom et al., 2016; Ruffell and Coussens, 2015). One of the limitations of this study is the lack of longitudinal biopsies, as only one tumor sample from each patient was analyzed. This limits the ability to perform clonality analysis using software and tools that are designed for multiple tumor samples. Extending this study to cohorts with multiple tumor samples would strengthen the results by allowing for more comprehensive analysis of clonal evolution over time.

Another limitation of this study is that the results regarding the association between phylogenies and their drivers are purely descriptive. To obtain a causal relationship between these drivers and tumor phylogenies, further experiments involving the genes GTF2IRD2B, TP53, and CTNNB1 and their relationship with clonal evolution are needed. To overcome these limitations, future studies could include longitudinal biopsies and analyze multiple tumor samples from each patient to gain a more comprehensive understanding of clonal evolution. Additionally, functional experiments could be conducted to explore the causal relationship between the identified drivers and tumor phylogenies. Such studies could ultimately enhance our understanding of the underlying mechanisms driving tumor progression and inform the development of more effective therapeutic strategies.

This study links the deep branching tree to the neutral evolution model, where no selection occurs, resulting in a tree with numerous clones with similar proportions. Therefore, this study corroborates the three well-known tumor evolution models, namely linear, branching, and neutral.

Several studies have shown a correlation between tumor heterogeneity and poor survival (Friemel et al., 2015; Roth et al., 2014). Our study confirms this association by demonstrating that high ITH is associated with poor prognosis in cancer since linear trees in our results have the lowest ITH, compared to the shallow and deep branching trees. For all cohorts, the phylogeny

type of the tree is associated with survival, so linear trees predict the best survival, and deep branching trees the worst survival.

Chapter 4: Comparing survival distributions of defined groups: Log-rank Tests

4.1 Log-rank test in survival analysis

Survival analysis is a statistical discipline that involves the examination of time-to-event data. Among the various statistical techniques available for analyzing survival data, the log-rank test is a popular method employed for comparing the survival distributions of multiple groups. Its widespread use in clinical trials and epidemiological studies, specifically in comparing the survival distributions of groups receiving different treatments, possessing diverse risk factors, or belonging to distinct populations, attests to its utility.

In addition to the standard log-rank test, there are several other variants that are commonly used in survival analysis. The weighted log-rank test is a modification of the standard log-rank test that assigns weights to different time periods in the survival analysis. By weighting the data, the weighted log-rank test can differently weight early versus late failures, while still comparing the survival distributions of the groups overall.

The stratified log-rank test is a modification of the standard log-rank test that takes into account the effect of one or more stratification variables. Stratification variables are used to identify strata within which the direction of the outcome parameter remains consistent, while accounting for potential differences in nuisance parameters across different strata. This approach allows for a more precise and reliable analysis of the outcome of interest in statistical models.

By stratifying the data, the stratified log-rank test can account for differences in survival between strata, while still comparing the survival distributions of the groups overall.

The covariate-adjusted log-rank test is another modification of the standard log-rank test that adjusts for the effects of one or more covariates. Covariates are variables that may influence survival times, but in the same way for different groups under the null hypothesis of no survival difference. By incorporating covariates, the log-rank test not only takes into account the impact of these variables on survival differences but also still allows for a comparison of overall survival distributions among different groups. This adjustment serves a crucial purpose in this work, as it has the potential to amplify the disparities in survival under the alternative hypothesis, without significantly influencing the null hypothesis.

4.1.1 Preliminaries

Notations. Let $\tilde{T}_i$ be death time and C_i be right censoring time for the i_{th} study subject. For a patient from a population under investigation, let the observed data be

$$(T_i, \delta_i, J_i, St_i, X_i), \quad i = 1, \ldots, n \tag{4.1}$$

where, for subject i, $T_i = \min(\tilde{T}_i, C_i)$ is the event (death) time, and $\delta_i = I(\tilde{T}_i \leq C_i)$ is the death indicator variable. Let J_i be a patient group indicator for $j = 1, 2, ..., K$. X_i is a p-dimensional vector containing possible survival-associated covariates that are not directly related to the patient group assignment variable. $St_i = l$ is the stratification label where L is the total number of strata. Denote the survival function with $S(t)$ and the the corresponding hazard function by $\lambda(t)$.

Counting processes. Define the following counting and at-risk processes:

$$N_{ijl}(t) = I(T_i \leq t, \delta_i = 1, J_i = j, St_i = l), \tag{4.2}$$

$$Y_{ijl}(t) = I(T_i \geq t, J_i = j, St_i = l). \tag{4.3}$$

The subscript $+$ notation indicates summation over the corresponding indices e.g. $N_{+j+} = \sum_{i=1}^{n} \sum_{l=1}^{L} N_{ijl}$. Denote $\pi_j = P(J_i = j)$, so $\sum_{j=1}^{K} \pi_j = 1$. Also n_j is the number of patients in outcome group j, so $\sum_{j=1}^{K} n_j = n$. As $n \to \infty$, $\frac{n_j}{n} \to \pi_j$ almost surely.

Stochastically ordered alternatives. The survival function S_1 is considered to be stochastically greater than another survival function S_2 if $S_1(t)$ is greater than or equal to $S_2(t)$ for all $t \geq 0$. This relationship is denoted as $S_1 \succeq S_2$. When additionally S_1 is strictly greater for some $t > 0$, then is denoted as $S_1 \succ S_2$ (Chang and McKeague, 2016). Consider the following null hypothesis for the equality of the following survival functions:

$$H_0 : S_1^l(t) = S_2^l(t) = ... = S_K^l(t), \quad l = 1, ..., L, \quad \forall t \geq 0, \tag{4.4}$$

against $H_1 - H_0$ where we consider H_1 as stochastically ordered alternative

$$H_1 : S_1^l \succ S_2^l \succ ... \succ S_K^l, \quad l = 1, ..., L. \tag{4.5}$$

where for some k, t we have $S_k(t) > S_{k+1}(t)$. This alternative hypothesis is particularly relevant in the context of different tumor phylogenies, where the aggressiveness of the tumor can be ordered, such as with branched tumor phylogenies being more aggressive than linear tumor

phylogenies. In such cases, we anticipate that more aggressive tumor types would exhibit lower survival rates.

Assumptions. We assume the following:

- (A0) Independence: $(T_i, \delta_i, J_i, St_i, X_i), i = 1, \ldots, n$, are independent identically distributed realizations of (T, δ, J, St, X).
- (A1) Non-informative censoring: $\tilde{T}_i$ and C_i are conditionally independent for $i = 1, ..., n$ given group indicator J_i, stratification label St_i, and covariates X_i.
- (A2) Under the null H_0 hypothesis, group indicator J_i is conditionally independent of covariates X_i given $Y_{i+l}(t) = 1$.

$$E(I_{[J_i=j]}|Y_{i+l} = 1, X_i = x) = E(I_{[J_i=j]}|Y_{i+l} = 1), \quad l = 1, 2, ..L. \tag{4.6}$$

- (A3) Stratification: Survival distributions are not the same in each stratum but the distribution of relevant covariates in each stratum is the same in each group.

4.1.2 The 2-sample log-rank test

The standard 2-sample log-rank test (Gehan, 1965; Mantel, 1966; Schmid et al., 1992; Therneau, 2023; Therneau and Grambsch, 2000) is well-known and well documented. There are no covariates or stratification variables in this case, but (4.2) and (4.3) notations can still be used for counting and at-risk processes where $L = 1$ so just for notational purposes denote $N_{ij1} = N_{ij+}$:

$$N_{ij+}(t) = I(T_i \leq t, \delta_i = 1, J_i = j), \quad j = 1, 2,$$

as counting processes and

$$Y_{ij+}(t) = I(T_i \geq t, J_i = j), \quad j = 1, 2$$

as at-risk processes.

Since we do not have St_i and X_i, the assumptions in Section 4.1.1 are simplified as follows:

- (A0) Independence: $(T_i, \delta_i, J_i), i = 1, \ldots, n$, are independent identically distributed realizations of (T, δ, J).
- (A1) Non-informative censoring: $\tilde{T_i}$ and C_i are conditionally independent for $i = 1, ..., n$ given group indicator J_i.

The test statistic. First, we define a statistic U as:

$$U = \int_0^{\tau_n} \{dN_{+1+}(t) - \frac{Y_{+1+}(t)}{Y_{+++}(t)} dN_{+++}(t)\} \tag{4.7}$$

$$= \int_0^{\tau_n} \left\{dN_{+1+}(t) - Y_{+1+}(t)\lambda(t)\,dt\right\} - \int_0^{\tau_n} \frac{Y_{+1+}(t)}{Y_{+++}(t)} \left\{dN_{+++}(t) - Y_{+++}(t)\lambda(t)\,dt\right\}$$

$$= \int_0^{\tau_n} \left\{dN_{+1+}(t) - Y_{+1+}(t)\lambda(t)\,dt\right\} - \int_0^{\tau_n} \frac{Y_{+1+}(t)}{Y_{+++}(t)} \left\{dN_{+1+}(t) - Y_{+1+}(t)\lambda(t)\,dt\right\}$$

$$- \int_0^{\tau_n} \frac{Y_{+1+}(t)}{Y_{+++}(t)} \left\{dN_{+2+}(t) - Y_{+2+}(t)\lambda(t)\,dt\right\}$$

$$= \int_0^{\tau_n} \frac{Y_{+2+}(t)}{Y_{+++}(t)} \left(dN_{+1+}(t) - Y_{+1+}(t)\,\lambda(t)\,dt\right) - \int_0^{\tau_n} \frac{Y_{+1+}(t)}{Y_{+++}(t)} \left(dN_{+2+}(t) - Y_{+2+}(t)\,\lambda(t)\,dt\right)$$

$$= \sum_{i=1}^{n} \left[\int_0^{\tau_n} \frac{Y_{+2+}(t)}{Y_{+++}(t)} \left(dN_{i1+}(t) - Y_{i1+}(t)\,\lambda(t)\,dt\right) - \int_0^{\tau_n} \frac{Y_{+1+}(t)}{Y_{+++}(t)} \left(dN_{i2+}(t) - Y_{i2+}(t)\,\lambda(t)\,dt\right)\right]$$

where bounded stopping time $\tau_n \leq \tau$ is defined so that $P(T_i \geq \tau_n) > 0$ for all i, but also such that the integral contains no time points where $Y_{+++}(t) < 2$. (τ is a point that satisfies $P(T_i \geq \tau) > 0$ for j = 1, 2). These stochastic integrals inside the summands in (4.7) can be viewed as being taken with respect to the following compensated counting process martingales (see Appendix).

$$M_{ij+}(s) \;=\; \int_0^{\min(s,\tau_n)} (dN_{ij+}(t) \;-\; Y_{ij+}(t)\,\lambda(t)\,dt).$$

Thus, we can rewrite the log-rank expression (4.7) as

$$U \;= \sum_{i=1}^{n} \int_0^{\tau_n} \Big[\frac{Y_{+1+}(t)}{Y_{+++}(t)}\,dM_{i2+}(t) \;-\; \frac{Y_{+2+}(t)}{Y_{+++}(t)}\,dM_{i1+}(t)\Big] \tag{4.8}$$

where the terms inside (4.9) both have mean 0 under null H_0. In certain large-sample settings related to martingale theory, variances can be estimated using the concept of "[predictable] variance" or cumulative conditional variance processes. The martingale central limit theorem (refer to Theorem 3 in the Appendix) and martingale convergence theorems provide a framework to establish the large-sample limits of normalized variances and demonstrate the large-sample consistency of variance and covariance estimators (Fleming and Harrington, 1991). By utilizing compensated counting-process martingales and stochastic-integral theory, expressions for variance can be derived (see Appendix, specifically Equation (A.2)). Under the null hypothesis H_0, the variance of the statistic U can be represented as:

$$\text{Var}(U) = \int_0^{\tau_n} E(\frac{Y_{+1+}(t)Y_{+2+}(t)}{Y_{+++}(t)})\lambda(t)dt. \tag{4.9}$$

It is noteworthy that the expressions inside the expectations in Equation (4.9) can be shown to converge in probability using the central limit theorems. Consequently, the normalized stochastic integral expressions converge in probability to their respective expectations. This convergence is facilitated by the fact that all the integral expressions are uniformly bounded and, as a result, uniformly integrable in this setting. Hence, the estimated variance is obtained, by substituting $d\hat{\Lambda}(t) = dN_{+++}(t)/Y_{+++}(t)$ in Equation (4.9)

$$\hat{\sigma}_U^2 = \int_0^{\tau_n} \frac{Y_{+1+}(t)Y_{+2+}(t)}{Y_{+++}^2(t)} dN_{+++}(t). \tag{4.10}$$

Finally, the 2-sample log-rank test statistic is given by:

$$Z = \frac{U}{\hat{\sigma}_U}. \tag{4.11}$$

Under the null hypothesis, the test statistic Z follows standard normal distribution. We can then calculate the asymptotic 2-sided p-value $2(1 - \Phi(|Z|)$ associated with our test statistic, which measures the large sample approximate probability of obtaining a test statistic as extreme or more extreme than the one we observed, assuming the null hypothesis is true.

4.1.3 The 2-sample stratified log-rank test

Suppose that we have two groups, as in log-rank test, but we want to control for a categorical covariate (e.g., type of a cancer (HCC, CCA)). Then there are 4 = 2 × 2 types of individuals.

Let $S_1^l(t)$ and $S_2^l(t)$ be the survival functions for stratum l of group 1 and group 2 respectively.

The null hypothesis is given as the following:

$$H_0 : S_1^l(t) = S_2^l(t), t > 0, l = 1, ..., L. \tag{4.12}$$

against which we consider the stochastically ordered alternative

$$H_1 : S_1^l(t) \succ S_2^l(t), t > 0, l = 1, ..., L \tag{4.13}$$

where the stochastic ordering is strict only in at least one of the strata.

The test statistic. First, we divide the data into L groups. Then, the 2-sample log-rank test statistic numerator is calculated for each group. The 2-sample stratified log rank test then is calculated as:

$$U_L = \sum_{l=1}^{L} \int_0^{\tau_n} \{dN_{+2l}(t) - \frac{Y_{+2l}(t)}{Y_{++l}(t)} dN_{++l}(t)\}. \tag{4.14}$$

Then we again rewrite U_L as

$$U_L = \sum_{l=1}^{L} \sum_{i=1}^{n} \int_0^{\tau_n} \left[\frac{Y_{+1l}(t)}{Y_{++l}(t)} \, dM_{i2l}(t) - \frac{Y_{+2l}(t)}{Y_{++l}(t)} \, dM_{i1l}(t) \right] \tag{4.15}$$

Because the elements in U_L are approximately uncorrelated, the estimated variance of U_L, using similar argument in Section 4.1.2 about the stochastic integrals and large sample theory, is given by:

$$\hat{\sigma}^2_{U_L} = \sum_{l=1}^{L} \int_0^{\tau_n} \frac{Y_{+1l}(t) Y_{+2l}(t)}{Y^2_{++l}(t)} dN_{++l}(t) \tag{4.16}$$

$$Z = \frac{U}{\hat{\sigma}_U} \tag{4.17}$$

Under the null hypothesis, the test statistic Z follows asymptotically for large samples standard normal distribution. We can then calculate the p-value associated with our test statistic, which measures the large sample approximate probability of obtaining a test statistic as extreme or more extreme than the one we observed, assuming the null hypothesis is true.

4.1.4 The K-sample covariate adjusted log-rank test

There may be additional covariates (such as age, gender, or diet) that are known or suspected to affect the survival outcomes (under alternatives to the null hypothesis) but also not directly related to patient group variable. The K-sample covariate adjusted log-rank test allows us to control for these covariates and examine the independent effect of the phylogeny-type variable on survival outcomes.

The test statistic Assume X is a p-dimensional observed baseline covariates to be adjusted in the construction of the test, with a nonsingular covariance matrix $\Sigma_X = Var(X)$. For fixed j, the ordinary log-rank statistic for distinguishing treatment group j versus group $J\backslash\{j\}$, which is already given in equation (4.9) for $K = 2$, can be written as the following:

$$U_{+j+} = \int_0^{\tau_n} \left\{ dN_{+j+}(t) \; - \; \frac{Y_{+j+}(t)}{Y_{+++}(t)} \, dN_{+++}(t) \right\} \; =$$

$$\int_0^{\tau_n} \left\{ dN_{+j+}(t) \; - \; Y_{+j+}(t)\lambda(t)\, dt \right\} \; - \; \int_0^{\tau_n} \frac{Y_{+j+}(t)}{Y_{+++}(t)} \left\{ dN_{+++}(t) \; - \; Y_{+++}(t)\lambda(t)\, dt \right\} \quad (4.18)$$

$$= \sum_{i=1}^{n} \Big[\int_0^{\tau_n} \big(1 - \frac{Y_{+j+}(t)}{Y_{+++}(t)}\big) \, (dN_{ij+}(t) - Y_{ij+}(t)\, \lambda(t)\, dt) - \sum_{k:k\neq j} \int_0^{\tau_n} \frac{Y_{+j+}(t)}{Y_{+++}(t)} \, (dN_{ik+}(t) - Y_{ik+}(t)\, \lambda(t)\, dt) \Big].$$

These stochastic integrals can be viewed as against the following compensated counting process martingales (Andersen and Gill, 1982)

$$M_{ik+}(s) \;=\; \int_0^{\min(s,\tau_n)} (dN_{ik+}(t) \;-\; Y_{ik+}(t)\,\lambda(t)\,dt).$$

Thus, we rewrite the log-rank expression (4.18)

$$U_{+j+} \;= \sum_{i=1}^{n} \int_0^{\tau_n} \Big[(1 - \frac{Y_{+j+}(t)}{Y_{+++}(t)})\, dM_{ij+}(t) \;- \sum_{k:k\neq j} \frac{Y_{+j+}(t)}{Y_{+++}(t)}\, dM_{ik+}(t)\Big] \;= \sum_{i=1}^{n} (O_{ij+}^{(1)} - O_{ij+}^{(2)}) \tag{4.19}$$

where $O_{ij+}^{(s)}$, s=1,2 are treated as mean 0 (under H_0) responses given $J_i = j$. Using martingale theory (see Appendix) adn similar argument we made in Section 4.1.2, we can easily show that the predictable variance is

$$\begin{aligned} \text{Var}(U_{+j+}) &= E\left(\int_0^{\tau_n} \left(1 - \frac{Y_{+j+}(t)}{Y_{+++}(t)}\right)^2 Y_{+j+}(t)\lambda(t)dt + \int_0^{\tau_n} \frac{Y_{+j+}^2(t)}{Y_{+++}^2(t)}(Y_{+++}(t) - Y_{+j+}(t))\lambda(t)dt\right) \\ &= \int_0^{\tau_n} E\left(Y_{+j+}(t) - \frac{Y_{+j+}^2(t)}{Y_{+++}(t)}\right)\lambda(t)dt. \end{aligned} \tag{4.20}$$

Again, martingale theory enables us to derive the covariances between U_{+j+} and $U_{+j'+}$, such that

$$\text{Cov}(U_{+j+}, U_{+j'+}) = -\int_0^{\tau_n} E\left(\frac{Y_{+j+}(t)Y_{+j'+}(t)}{Y_{+++}(t)}\right)\lambda(t)dt. \tag{4.21}$$

Ye et al. (2023) developed the adjusted log-rank statistic for the 2-sample case. In this section, we will extend their theory to the case $K > 2$ by utilizing the martingale theory to calculate the variances and covariances easily. Following their idea, the next step is to linearly project $O_{ij+}^{(1)}$

onto $I_{[J_i=j]} \cdot X_i$ and $O^{(2)}_{ij+}$ onto $I_{[J_i \neq j]} \cdot X_i$ covariates by defining

$$\beta^{(1)}_{j+} = (\text{Var}(X_i \mid J_i = j))^{-1}\text{Cov}(X_i,\, O^{(1)}_{ij+} \mid J_i = j) \tag{4.22}$$

$$\beta^{(2)}_{j+} = (\text{Var}(X_i \mid J_i \neq j))^{-1}\text{Cov}(X_i,\, O^{(2)}_{ij+} \mid J_i \neq j). \tag{4.23}$$

Let $\mu_X = E(X_i)$, $\mu_{X_j} = E(X_i|J_i = j)$, $\mu_{X_{-j}} = E(X_i|J_i = j)$, and $\bar{X}$ is the sample mean of all X_i's, $\bar{X}_j$ is the sample mean of X_i's with $J_i = j$, and $\bar{X}_{-j}$ is the sample mean of X_i's with $J_i \neq j$. Thus, the projections

$$P^{(1)}_{ij+} = O^{(1)}_{ij+} - I_{[J_i=j]}(X_i - \mu_{X_j})^T\beta^{(1)}_{j+} \tag{4.24}$$

$$P^{(2)}_{ij+} = O^{(2)}_{ij+} - I_{[J_i\neq j]}(X_i - \mu_{X_{-j}})^T\beta^{(2)}_{j+} \tag{4.25}$$

have means 0 under the null-hypothesis assumption that X_i are iid given $J_i = j$ over all i, j. Then, we have the adjusted version of U_{+j+}

$$U^{(adj)}_{+j+} = \frac{1}{n}\sum_{i=1}^{n}(P^{(1)}_{ij+} - P^{(2)}_{ij+}) \tag{4.26}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(O^{(1)}_{ij+} - O^{(2)}_{ij+}) - \frac{1}{n}\sum_{i=1}^{n}\{I_{[J_i=j]}\,(X_i - \mu_{X_j})^T\beta^{(1)}_{j+} - \, I_{[J_i\neq j]}\,(X_i - \mu_{X_{-j}})^T\beta^{(2)}_{j+}\}$$

where expectation of the last term is 0 since the covariate X_i is independent of group outcome J_i under null hypothesis as a consequence of assumption (A2) at $t = 0$.

Since the λ, β, and μ parameters are unknown in the adjusted expression (4.26), the actual feasible K-sample statistic entries must involve estimates substituted for them. First, we substitute

$d\hat{\Lambda}(t) = dN_{+++}(t)/Y_{+++}(t)$ and we obtain

$$\hat{O}_{ij+}^{(1)} = \int_0^{\tau_n} \left(1 - \frac{Y_{+j+}(t)}{Y_{+++}(t)}\right) \left(dN_{ij+}(t) - \frac{Y_{ij+}(t)}{Y_{+++}(t)}\, dN_{+++}(t)\right)$$

$$= \int_0^{\tau_n} \left(1 - \frac{Y_{+j+}(t)}{Y_{+++}(t)}\right) \left(dM_{ij+}(t) - \frac{Y_{ij+}(t)}{Y_{+++}(t)}\, dM_{+++}(t)\right)$$

and similarly

$$\hat{O}_{ij+}^{(2)} = \sum_{k:\, k\neq j} \int_0^{\tau_n} \frac{Y_{+j+}(t)}{Y_{+++}(t)} \left(dN_{ik+}(t) - \frac{Y_{ik+}(t)}{Y_{+++}(t)}\, dN_{+++}(t)\right)$$

$$= \sum_{k:\, k\neq j} \int_0^{\tau_n} \frac{Y_{+j+}(t)}{Y_{+++}(t)} \left(dM_{ik+}(t) - \frac{Y_{ik+}(t)}{Y_{+++}(t)}\, dM_{+++}(t)\right).$$

Next, we argue two approaches to substitute μ_{X_j} and $\mu_{X_{-j}}$ parameters. First, to substitute them by $\bar{X}_j$ and $\bar{X}_{-j}$, and get

$$\hat{\beta}_{j+}^{(1)} = \Big[\sum_{i:J_i=j} (X_i - \bar{X}_j)^{\otimes 2}\Big]^{-1} \sum_{i:J_i=j} (X_i - \bar{X}_j)\, \hat{O}_{ij+}^{(1)} \tag{4.27}$$

and

$$\hat{\beta}_{j+}^{(2)} = \Big[\sum_{i:J_i\neq j} (X_i - \bar{X}_{-j})^{\otimes 2}\Big]^{-1} \sum_{i:J_i\neq j} (X_i - \bar{X}_{-j})\, \hat{O}_{ij+}^{(2)}. \tag{4.28}$$

here, in the case where $K = 2$, formulas (4.27) and (4.28) do agree with the formulas given by Ye et al. (2023), however; both their and our proof for the main theorem utilize the expression when we substitute those parameters by $\bar{X}$ instead as second approach. Therefore, we suggest to

use $\bar{X}$ to be able to follow their proof strategy. Thus, the estimator we use are the following:

$$\hat{\beta}_{j+}^{(1)} = \Big[\sum_{i:J_i=j} (X_i - \bar{X})^{\otimes 2} \Big]^{-1} \sum_{i:J_i=j} (X_i - \bar{X})\, \hat{O}_{ij+}^{(1)} \tag{4.29}$$

and

$$\hat{\beta}_{j+}^{(2)} = \Big[\sum_{i:J_i\neq j} (X_i - \bar{X})^{\otimes 2} \Big]^{-1} \sum_{i:J_i\neq j} (X_i - \bar{X})\, \hat{O}_{ij+}^{(2)}. \tag{4.30}$$

Also, when we substitute all the parameters with their aforementioned estimators,

$$\hat{P}_{ij+}^{(1)} = \hat{O}_{ij+}^{(1)} - I_{[J_i=j]}\, (X_i - \bar{X})^T \hat{\beta}_{j+}^{(1)} \tag{4.31}$$

$$\hat{P}_{ij+}^{(2)} = \hat{O}_{ij+}^{(2)} - I_{[J_i\neq j]}\, (X_i - \bar{X})^T \hat{\beta}_{j+}^{(2)}, \tag{4.32}$$

we get the following adjusted statistic to compare the treatment group j versus group $J\backslash\{j\}$

$$\hat{U}_{+j+}^{(adj)} = \frac{1}{n} \sum_{i=1}^{n} (\hat{P}_{ij+}^{(1)} - \hat{P}_{ij+}^{(2)}). \tag{4.33}$$

Denote $\theta_j^{(1)} = E(O_{ij+}^{(1)} | J_i = j)$ and $\theta_j^{(2)} = E(O_{ij+}^{(2)} | J_i \neq j)$. We know under the null hypothesis that $\theta_j^{(1)} = \theta_j^{(2)} = 0$.

Next, we present a lemma that investigates the asymptotic behavior of the estimators for the β parameters, which will be instrumental in proving Theorem 1. The proof of this lemma relies on the application of the laws of large numbers, the assumption of independent and identically distributed (iid) observations, and the validity of assumption (A2). The proof method resembles to the proof outlined in Ye et al. (2023).

Lemma 1 *Under assumption (A2), estimators 4.27 and 4.28 converge to parameters 4.22 and*

4.23 in probability:

$$(a) \quad \hat{\beta}_{j+}^{(1)} = \beta_{j+}^{(1)} + o_p(1)$$

$$(b) \quad \hat{\beta}_{j+}^{(2)} = \beta_{j+}^{(2)} + o_p(1)$$

.

Theorem 1, which we present next, extends and builds upon the theory developed in Ye et al. (2023). The proof of this theorem is provided in the Appendix. Specifically focused on the null hypothesis, the theorem offers important information about the behavior of our proposed K-sample covariate-adjusted log-rank test.

Theorem 1 *Assume (A0)-(A2). Then, the following results hold.*

(a) Under the null H_0 hypothesis

$$\sqrt{n}\hat{U}_{+j+}^{(adj)} \to \mathcal{N}(0, \sigma_{adj}^2)$$

in distribution, where

$$\sigma_{adj}^2 = \sigma^2 - \pi_j(1-\pi_j)(\beta_{j+}^{(1)} + \beta_{j+}^{(2)})^T \Sigma_X (\beta_{j+}^{(1)} + \beta_{j+}^{(2)})$$

and

$$\sigma^2 = \pi_j Var(O_{ij+}^{(1)}) + (1-\pi_j) Var(O_{ij+}^{(2)})$$

(b) Under the null H_0 hypothesis, $\hat{\sigma}^2_{adj} \xrightarrow{p} \sigma^2_{adj}$ and $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$ where

$$\hat{\sigma}^2_{adj} = \hat{\sigma}^2 - \pi_j(1-\pi_j)(\hat{\beta}^{(1)}_{j+} + \hat{\beta}^{(2)}_{j+})^T \hat{\Sigma}_X (\hat{\beta}^{(1)}_{j+} + \hat{\beta}^{(2)}_{j+}),$$

and

$$\hat{\sigma}^2 = \int_0^{\tau_n} Y_{+j+}(t) \left(1 - \frac{Y_{+j+}(t)}{Y_{+++}(t)}\right) \frac{dN_{+++}(t)}{Y_{+++}(t)}.$$

Finally, to construct our K-sample covariate adjusted log-rank test, we will utilize the quadratic form of the estimated adjusted statistics (K-1 of them will be utilized). From Theorem 1, we know under the null,

$$\hat{U}^{(adj)}_{+j+} \to \mathcal{N}(0, \hat{\sigma}^2_{adj}).$$

And the K-sample covariate-adjusted log-rank statistic is given as

$$Q = (\hat{U}^{(adj)}_{+1+}, \hat{U}^{(adj)}_{+2+}, ..., \hat{U}^{(adj)}_{+(K-1)+})^T \hat{\Sigma}^{-1} (\hat{U}^{(adj)}_{+1+}, \hat{U}^{(adj)}_{+2+}, ..., \hat{U}^{(adj)}_{+(K-1)+}) \tag{4.34}$$

which we believe that the limiting distribution of Q will have a chi-square distribution with $(K-1)$ degrees of freedom under the null hypothesis. There is a probability-limit invertible matrix Σ which is consistently estimated by $\hat{\Sigma} = (\hat{\Sigma}_{jj'})_{(K-1)x(K-1)}$ whose elements are given in the following

$$\hat{\Sigma}_{jj} = \int_0^{\tau_n} \left(\frac{Y_{+j+}}{Y_{+++}}(t) - \frac{Y^2_{+j+}(t)}{Y^2_{+++}(t)}\right) dN_{+++}(t) - \pi_j(1-\pi_j)(\hat{\beta}^{(1)}_{j+} + \hat{\beta}^{(2)}_{j+})^T \hat{\Sigma}_X (\hat{\beta}^{(1)}_{j+} + \hat{\beta}^{(2)}_{j+}),$$

and

$$\hat{\Sigma}_{jj'} = -\int_0^{\tau_n} \left(\frac{Y_{+j+}(t) Y_{+j'+}(t)}{Y^2_{+++}(t)} \right) dN_{+++}(t) + \pi_j \pi_{j'} (\hat{\beta}^{(1)}_{j+} - \hat{\beta}^{(2)}_{j+})^T \hat{\Sigma}_X (\hat{\beta}^{(1)}_{j'+} - \hat{\beta}^{(2)}_{j'+})$$

where $j \neq j'$, and $\hat{\Sigma}_X$ is the estimated covariance matrix for all X_i variables. Furthermore, we substitute $\pi_j = \frac{n_j}{n}$ and $\pi_j = \frac{n_{j'}}{n}$. The calculations for the covariance terms are provided in the Appendix.

4.1.5 The K-sample stratified covariate-adjusted log-rank test

To perform the stratified adjusted log-rank test, the data is divided into L groups, and the test statistic in Equation (4.26) is calculated for each group. The elements of the adjusted stratified log-rank test is given by:

$$U^{(adj)}_{L,+j+} = \sum l = 1^L \frac{1}{n} \sum_{i=1}^{n} (P^{(1)}_{ijl} - P^{(2)}_{ijl}), \tag{4.35}$$

where $P^{(1)}_{ijl}$ and $P^{(2)}_{ijl}$ represent the projections in the l^{th} stratum for the comparison of group j versus group $J \backslash j$.

Next, the calculations for estimating variances and covariances are performed within each group. Since the elements in $U^{(adj)}_{L,+j+}$ are approximately uncorrelated, the sum of the estimated covariance terms is used to calculate the K-sample stratified covariate-adjusted log-rank test.

4.2 Validation of constructed tumor evolution tree clusters using survival analysis

We want to investigate to what extent covariates influence the survival curves of already constructed tumor phylogeny clusters. Our objective is twofold:

Validation. We seek to determine the validity and meaningfulness of the constructed clusters by examining their association with survival outcomes, considering additional information. To achieve this, we employ a stratified log-rank test, which enables us to assess whether there are statistically significant differences in survival curves between the tree clusters. If the test reveals a significant difference in survival between the clusters, it provides evidence in support of the validity of the grouping. Conversely, if the survival curves do not significantly differ, it suggests that the clusters may not represent distinct biological groups.

Detection. Additionally, we aim to investigate the impact of the given information on the differences observed in survival curves. To assess the influence of additional covariates on the separation of the tree variable in a survival analysis, we employ the covariate-adjusted log-rank test. This test allows us to determine whether the inclusion of specific covariates improves the discrimination between the tree clusters in terms of survival outcomes. By analyzing the significance of the covariate-adjusted log-rank test, we gain understanding about whether the provided information contributes to the observed differences in survival curves.

Through these validation and detection analyses, we can better understand the robustness and significance of the constructed tumor evolution tree clusters, as well as the impact of additional covariates on survival outcomes.

4.2.1 Application to real data

In Chapter 1, we introduce three liver cancer cohorts and TIGER-LC (Chaisaingmongkol et al., 2017) is one of them. Although we have only used Hepatocellular Carcinoma (HCC) patients from that cohort for the previous analyses, there are also Cholangiocarcinoma (CCA) patients followed up in this cohort. (HCC) and (CCA) are clinically distinct primary liver cancers with etiological and biological heterogeneity. To control for the effect of cancer sub-type while comparing the survival distributions of tumor evolution trees with log-rank test, we use cancer sub-type as a stratification variable in our model. TIGER-LC data also includes some etiological, demographic and clinical features such as, age, gender, tumor type that may affect survival outcomes.

The censoring reason for the subjects in TIGER-LC cohort is either the end-of-study events or instances of dropout with unknown reasons. For the given dataset, the lifetime variable $\tilde{T}$ represents the time from the first diagnosis until death.

Variables	**Count**	**Censoring %**
Phylogeny Type		
Linear	75	41%
Shallow Branching	104	31%
Deep Branching	70	21%
Cancer Stage		
Early	39	49%
Late	35	6%
Missing	175	
Cancer Type		
CCA	171	22%
HCC	78	49%
Age		
Old	105	32%
Young	132	30%
Missing	12	
Gender		
Female	81	24%
Male	158	34%

Table 4.1: Data Summary for TIGER-LC cohort (n=249)
Source: Chaisaingmongkol et al., 2017.

Table 4.1 displays a descriptive summary of survival data for TIGER-LC cohort. Here, the outcome variable is tumor phylogeny type (Table 4.1) with 3 levels, namely linear with sample size 75, shallow branching with sample size 104, and deep branching with sample size 70, and the stratification variable we would like to use is the cancer type with 78 HCC and 171 CCA patients. Covariates we considered to adjust for are age, gender and cancer stage. Before we delve into the data structure, it is crucial to acknowledge a significant limitation, the small sizes of the groups. In fact, they are tiny, which poses a severe constraint on the statistical power and generalizability of our findings. Given these limitations, it is important to present the adjusted and stratified analyses based on the TIGER-LC dataset as purely ”illustrative” rather

than scientifically definitive. While the results offer valuable information about the application of our methodology, it is essential to recognize that the small group sizes may limit the correctness of the asymptotic null-hypothesis reference distribution of the statistics and the robustness of the conclusions drawn from this particular study.

When analyzing this survival dataset, we are interested in estimating the probability of survival over time for each phylogeny group. However, the survival probability may differ between phylogeny groups due to differences in these baseline covariates. Ignoring these differences can lead to biased estimates and incorrect conclusions. Therefore, we further adjust for these covariates while performing log-rank test to compare survival distributions of tumor evolution trees or phylogeny groups.

In Figure 4.1 (a), we present the survival curves without any sort of stratification, and the corresponding 3-sample log-rank p-value is reported as an overall hypothesis test for equality of all survival curves. In Figure 4.1 (b), we show the survival curves stratified by cancer type. The corresponding 3-sample stratified log-rank p-values are reported. From the results, we conclude tentatively that when stratifying by cancer type, the survival distributions differ among the strata, indicating a significant association between tumor phylogeny and survival time for at least one cancer type.

Similarly, in Figure 4.2 (a), we present the survival curves for all patients from three cohorts without stratification. This may allow us to assess the overall differences in survival among the tumor phylogenies, regardless of the specific cohort they belong to.

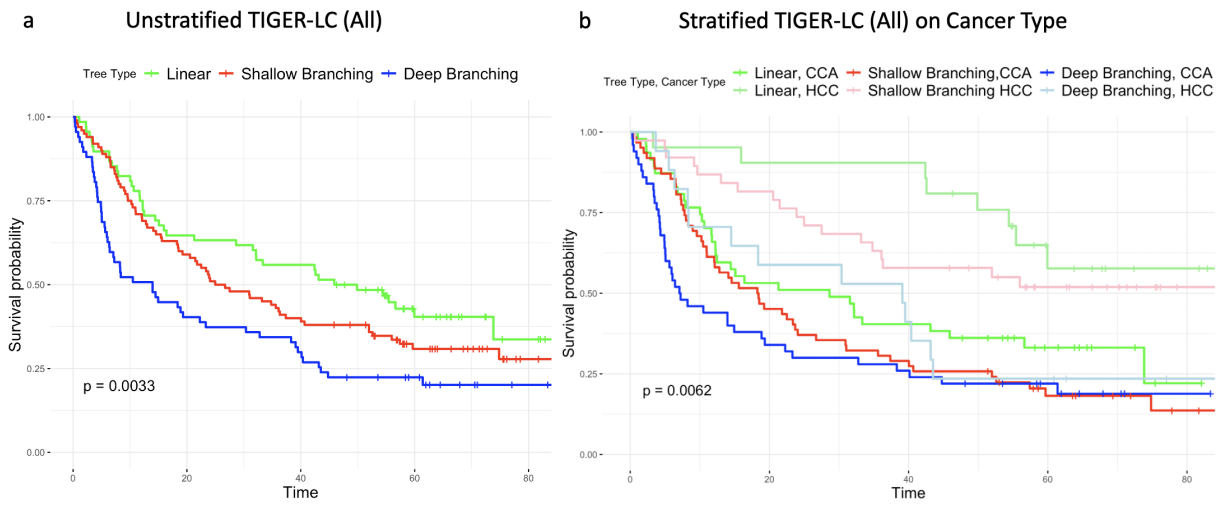


Figure 4.1: Survival curves for TIGER-LC (a) without stratification and (b) stratified by cancer type. Missing values for covariates were included in the construction of Kaplan-Meier curves.

It is important to note that our analysis was conducted separately on each cohort. We chose to analyze each cohort separately rather than applying stratification to the combined dataset in this study. By adopting this approach, we are able to capture the distinct characteristics and survival patterns within each cohort individually. This not only allows for a comprehensive analysis of each cohort's data but also serves as an additional validation method for the constructed tree clusters. By examining the survival trends within each cohort separately, we can assess the consistency of the results across cohorts and gain a more robust understanding of the underlying patterns. In Figure 4.2 (b), the survival curves are stratified by cohort, providing understanding into the differences in survival within each cohort.

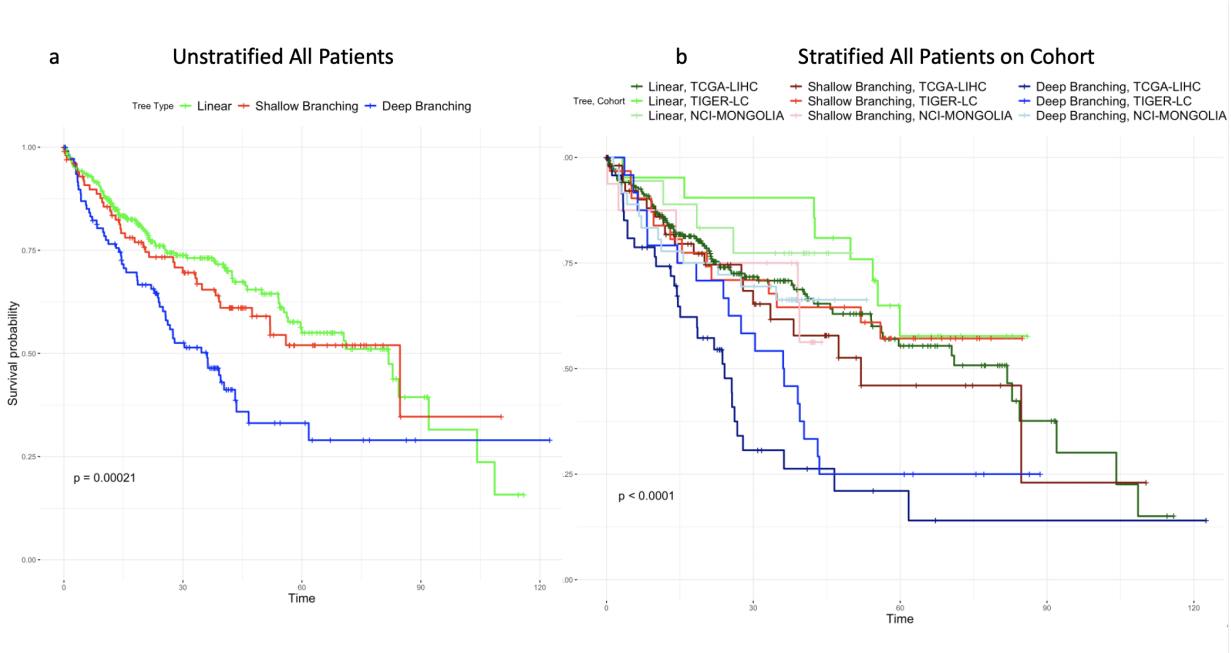


Figure 4.2: Survival curves for all patients (a) without stratification and (b) stratified on cohort. Missing values for covariates were included in the construction of Kaplan-Meier curves.

Next, we perform an adjustment for age and gender variables in the TIGER-LC cohort analysis. We calculate the adjusted test statistic by incorporating age and gender as covariates. Table 4.2 presents the test statistics and their corresponding p-values for the different log-rank tests conducted. Here, the p-values are calculated using the Chi-square distribution table based on the statistic Q in Equation 4.34. The unadjusted log-rank test corresponds to the 3-sample log-rank test or, equivalently, the score test of a Cox model with only the ”tree” variable included. The adjusted log-rank test involves adjusting for age and gender, while the stratified log-rank test stratifies the analysis based on the cancer type variable. From the table, it is observed that adjusting for gender and age yields a more significant test statistic compared to the unadjusted test.

	Test	Test Statistic	P Value
1	Unadjusted 3-Sample Log-Rank	11.45	0.0033
2	Adjusted 3-Sample Log-Rank	102.33	0.0000
3	Stratified 3-Sample Log-Rank	10.17	0.0062
4	Adjusted Stratified 3-Sample Log-Rank	99.99	0.0000

Table 4.2: Adjustment covariates: Age and Gender, Stratification variable: Cancer type. All 4 tests are applied after excluding all missing values for age and gender (n=235). Test Statistics and P-values are calculated using formulas in Section 4.1 for TIGER-LC cohort.

Despite a substantial reduction in the sample size due to missing values in the cancer stage variable, we conducted another adjustment for stage and gender variables purely for illustrative purposes. This adjustment involved calculating the adjusted test statistic by incorporating stage and gender as covariates. The results of the different log-rank tests performed, along with their corresponding p-values, are presented in Table 4.3. From the results, we observe that adjusting for gender and stage yields significant test statistic compared to the unadjusted test.

	Test	Test Statistic	P Value
1	Unadjusted 3-Sample Log-Rank	5.84	0.0540
2	Adjusted 3-Sample Log-Rank	17.22	0.0001
3	Stratified 3-Sample Log-Rank	5.53	0.0630
4	Adjusted Stratified 3-Sample Log-Rank	20.56	0.0000

Table 4.3: Adjustment covariates: Stage and Gender, Stratification variable: Cancer type. All 4 tests are applied after excluding all missing values (n=73). Test Statistics and P-values are calculated using formulas in Section 4.1 for TIGER-LC cohort.

4.3 Discussion

In this chapter, our primary objective was to propose the application of log-rank tests for validating cluster differences in survival analysis. The log-rank test is a well-established statistical method used to compare survival distributions between groups, making it suitable for

our investigation.

We began by discussing the theoretical foundations and technicalities of the log-rank test. However, the key focus of our discussion lies in utilizing the log-rank test for cluster validation. To ensure the robustness of our analysis, we introduced Assumption (A2), in our specific problem it is regarding the conditional independence of the phylogeny indicator and the baseline covariates that are adjusted in the log-rank test. This assumption is plausible because while genomics and phylogeny play a significant role in cancer research and survival, there are several covariates that can also impact cancer outcomes but are not directly related to genomics or its inference outcome, phylogeny. For instance, socioeconomic status emerged as an important covariate impacting cancer outcomes. Factors such as income, education level, and access to healthcare can contribute to disparities in survival rates. Considering all these covariates, we stress the importance of validating cluster differences in survival analysis beyond solely focusing on genomics. By accounting for a broader range of factors, we enhance our understanding of the complexities underlying phylogeny cluster disparities and their impact on survival outcomes. However, it is essential to recognize that our data have some limitations concerning these covariates, and there is a possibility that the assumption (A2) may be violated. In Section 5.1, we discuss in detail the limitations of our covariate choices given the available data and the potential challenges in satisfying this assumption. Despite these limitations, we proceed with the illustrative use of the K-sample adjusted log-rank test on our data to shed light on the importance of considering covariates in survival analysis. This discussion further strengthens the scientific rigor of our research and guides future investigations in the survival analysis and cancer research.

In this study, it is essential to recognize certain critical issues that emerged during the analysis. These included the presence of tiny outcome groups within strata, significant censoring

in the data, and the use of inappropriate adjustment covariates. As a consequence of these challenges, the primary purpose of our analysis is purely illustrative, aim at demonstrating the application of the developed methodology. While we acknowledge these limitations, it is crucial to emphasize that the study served as an opportunity to showcase the potential of the new method.

Overall, Chapter 4 demonstrated the utility of the log-rank test for comparing survival distributions between defined groups, specifically in the context of cluster validation. By incorporating covariates beyond genomics and phylogeny, we gain a more comprehensive perspective on the factors influencing cluster differences in survival outcomes.

Chapter 5: Conclusions

5.1 Implications and limitations

In this research, our primary focus has been to explore tumor clonality and its implications. While previous works (Castelli et al., 2017; Davis et al., 2017; Vendramin et al., 2021; Zhu et al., 2021) have contributed significantly to our understanding of tumor clonality and evolutionary trajectories across various cancers, a noteworthy gap in the literature lies in the lack of comparative analysis using real-world data. To address this gap, we conducted our analysis on three independent liver cancer cohorts. For each cohort, we represented tumor cell lineages as trees and identified distinct clonal tree clusters, including linear, shallow branching, and deep branching trees. These identified clusters exhibited unique characteristics, with linear clusters displaying higher immune activity and less aggressiveness on average, while deep branching clusters showed the most aggressive behavior. By conducting this comprehensive analysis across multiple cohorts, we aimed to shed light on the diverse consequences of tumor clonality in liver cancer.

One recent article focused on Malignant Pleural Mesothelioma (MPM) and examined tumor evolution models in the context of linear and branching tree clusters (Zhang et al., 2021). While this study provided understanding into MPM's clonality, it only explored linear versus branching patterns. Unlike this study's focus on linear and branching trees, we extended the analysis to include additional clusters, namely deep branching and shallow branching tree patterns.

Also, Zhang et al. (2021) did not specifically compare the survival outcomes between linear and branching tree clusters. In contrast, our research extended beyond the exploration of clonal tree patterns and delved into an analysis of survival outcomes in liver cancer. By specifically comparing these tree clusters, we aimed to uncover potential differences in patient prognoses and survival trajectories.

In alignment with the previous study by Zhang et al. (2021) that explored the tumor microenvironment, our research also delved into the intricate relationship between the clonal tree structures and the tumor microenvironment in liver cancer. However, instead of comparing tumor microenvironment cells for high clonality versus low clonality, as Zhang et al. (2021) did, we took a different approach in our analysis. Specifically, we focused on comparing the tumor microenvironment cells associated with different clonal tree structures, namely linear, deep branching, and shallow branching tree clusters. By examining these distinct clonal tree patterns, we aimed to elucidate how each type of clonality relates to the composition and characteristics of the tumor microenvironment.

The consequences of clonality in liver cancer are diverse and significantly impact disease progression and clinical outcomes. Some key consequences are as follows:

Aggressiveness: Clonality influences the aggressiveness of liver cancer, with deep branching clusters displaying the most aggressive behavior. These clusters exhibit rapid growth, invasiveness, and a higher likelihood of metastasis, contributing to a more challenging clinical course.

Treatment Failure: The presence of distinct clonal populations within a tumor can lead to treatment failure. Different clones may respond differently to therapies, resulting in the survival of treatment-resistant subclones, leading to disease recurrence and progression.

Immune Response: Clonal heterogeneity influences the immune response within the tumor microenvironment. Linear clusters, being more immune active, may be more susceptible to

immune-mediated clearance.

Diagnosis Challenges: Tumor clonality poses challenges in accurate diagnosis. The presence of genetically diverse subclones may lead to mischaracterization of the tumor's aggressiveness and the potential underestimation of its clinical behavior.

Prognosis: Clonality has significant prognostic implications. Patients with deep branching clusters generally have a worse prognosis due to their aggressive nature and increased likelihood of treatment resistance and metastasis.

In conclusion, my research on dissecting tumor clonality in liver cancer and representing tumors as trees has shed light on the different clonal clusters and their implications. The identified linear, shallow branching, and deep branching clusters have revealed significant differences in terms of immune activity and aggressiveness.

Our research also emphasizes cluster validation by employing the covariate-adjusted log-rank test. The process of selecting appropriate covariates demands careful consideration, ensuring that they not only satisfy the assumptions in our covariate-adjusted log-rank theory but are also associated with survival outcomes under the alternative hypotheses.

In this study, we extended the 2-sample covariate-adjusted log-rank test to a K-sample covariate-adjusted test. It is important to note that this methodological extension is a valuable contribution to the field, allowing for a more comprehensive analysis of survival differences among different outcome groups. However, it is equally essential to make cautious and supportable claims about the findings derived from our data analysis in the specific context of tumor phylogeny types. While our approach offers valuable insights and sheds light on the potential influence of covariates, the interpretability of the results should be approached with care, given the potential limitations of small group sizes and other factors that may affect the robustness of the conclusions.

As such, further research on larger datasets is warranted to validate and strengthen the implications of our findings in the domain of tumor phylogeny types.

In Section 4.2.1, we applied this adjustment to TIGER-LC cohort where the covariates adjusted for are the stage of the cancer and the gender of the patient. While adjusting for covariates is essential to obtain more accurate estimates and improve the statistical power of the analysis, it is important to acknowledge potential challenges and limitations associated with the chosen covariates. One notable concern is the potential violation of assumption (A2) in 4.1.1, which states that under the null hypothesis, the outcome group variable is conditionally independent of the covariates given that the patient is at risk. In the context of the data, this assumption implies that the stage of the cancer or the age of the patient, as a covariates, are not causally related to the phylogeny group, the outcome of interest. However, it is plausible that the stage of the cancer could be related to the phylogeny group which is an inferred outcome from genomics data. Age at disease occurrence can also often be related to genomics in the context of cancer research. Genomic alterations and mutations in cells can accumulate over time, and age-related changes in the genome are often associated with an increased risk of developing certain types of cancer. Therefore, considering the inferred outcome from genomics data and the relationship between age and genomic changes in cancer, this assumption may be at risk of being violated, leading to biased results.

Furthermore, while the choice of covariates in this study involves adjusting for stage, age, and gender, there may be other potential confounders or relevant covariates that were not included in the analysis. For instance, other clinical or molecular factors related to the cancer's biology might impact the phylogeny group and should be considered in future research.

In conclusion, the extension of the covariate-adjusted log-rank test to a K-sample test

represents a valuable contribution to survival analysis. However, in application of this method to genomic cancer datasets, the choice of covariates warrants careful consideration due to its potential relationship with the outcome variable. By acknowledging and addressing these complexities, researchers can ensure the robustness and scientific validity of their findings, and pave the way for more accurate and informative analyses in the field of survival analysis and cancer research.

5.2 Future work

There are several avenues for future investigation that will allow us to build upon the findings presented here and contribute further to the field of genomic analysis and cancer research.

Chapter 2 of this thesis focuses on comparing copy number comparison tools using real data from liver cancer. While the results obtained from this analysis provided some information into the performance of these tools, it is prudent to further investigate their effectiveness using real and simulated data. Simulating a complex data set of sequencing data poses a significant challenge, but it is a desirable step toward evaluating the tools' performance under controlled conditions. Conducting a comprehensive analysis using simulated data sets that closely resemble the characteristics of real sequencing data would be one of the avenues toward improvements of the methods in this thesis. This will involve developing realistic simulation models that capture the complexities of genomic variations, including copy number alterations, in liver cancer. By carefully designing the simulated data sets, we can systematically evaluate the performance of the copy number comparison tools and gain a better understanding of their strengths and limitations.

Chapter 3 of this thesis represents our work that has been submitted to a journal. To enhance the quality and impact of our research, we plan to undertake additional revisions based

on valuable feedback received during the peer review process. Specifically, we plan to conduct additional analyses by considering additional variables for adjustment and stratification, as discussed in Chapter 4. This will improve the robustness of our findings.

In Chapter 4, we identified several promising directions for future research and potential manuscripts. One key aspect that warrants further exploration is the behavior of test statistics under alternative hypotheses in the context of adjusted log-rank testing. Although we followed the approach proposed by Ye et al. (2023), there is an opportunity to thoroughly investigate the performance of these statistics when faced with different underlying assumptions and scenarios. By conducting comprehensive simulations and theoretical analyses, we can gain a deeper understanding of the statistical properties and robustness of our proposed methodology. The outcomes of this investigation will be crucial in establishing the applicability and generalizability of our approach across diverse datasets and clinical contexts. We anticipate that the results of this study will serve as a solid foundation for a manuscript dedicated to the theoretical foundations and practical implications of adjusted log-rank testing for the purpose of cluster validation that were defined using genomics. Additionally, we intend to expand the applicability of our methodology to different types of genomic datasets,forms of clustering using genomic data other than phylogenetic representation. Our goal is to strengthen the conclusions with the theoretical foundations of the log-rank test.

As with any research endeavor, this thesis has certain limitations that open avenues for future work. One such limitation is the absence of multiple samples from a patient in the available dataset. Obtaining and incorporating additional data with multiple samples from patients would enable a more comprehensive analysis of tumor evolution and clonality. This would allow us to capture a broader range of genetic variations and better characterize the subclonal dynamics

within tumors. By leveraging datasets with longitudinal samples, we can investigate the temporal evolution of tumors and elucidate the mechanisms driving their progression. Integrating multi-sample data into our analysis will require the development of innovative statistical models and computational tools to handle the increased complexity and heterogeneity of the data. The outcomes of this future research endeavor will be instrumental in refining our understanding of tumor evolution dynamics and improving the accuracy of our analyses.

In summary, this thesis has laid the groundwork for significant contributions. As we move forward, we will conduct additional revisions to enhance the quality of our work, thoroughly investigate the behavior of test statistics under alternative hypotheses, expand the applicability of our methodology to diverse genomic datasets, and address the limitations of the current study. These future research directions will not only strengthen our conclusions but also provide valuable insights into tumor evolution, guide clinical decision-making, and pave the way for improved diagnostic and therapeutic strategies in the field of cancer research.

Appendix A: Theory and Proofs

A.1 On martingale theory

We present the following definitions and theorems to establish a connection between the statistics we discus and the concepts and known theoretical results concerning martingales (Fleming and Harrington, 1991; Rogers and Williams, 1994). First, we note that all filtrations in our settings are right-continuous, and all processes cadlag, i.e. assumed to be a.s. right-continuous with limits from the left at every point. Left-continuous processes (such as the at-risk processes) are special cases of a more general technical concept called "predictability".

Martingale. Suppose $X(t)$ is a right-continuous stochastic process with left-hand limits that is adapted to filtration $(\mathcal{F}_t)$. $X(t)$ is a martingale if:

1. $E|X(t)| < \infty$ for all t.

2. $E[X(t+s) \mid \mathcal{F}_t] = X(t)$ almost surely for all $t \geq 0$ and $s \geq 0$.

The idea behind a martingale is that given the information available at time t (represented by $\mathcal{F}_t$), the expected value of $X(t+s)$ is equal to $X(t)$. Moreover, $X(t)$ is called a sub-martingale if the inequality in condition (b) is replaced by $\leq$, and it is called a super-martingale if the inequality in condition (b) is replaced by $\geq$.

We also note that every counting process is a submartingale and the compensator process

$A(t) = \int_0^t Y(s)\lambda(s)ds$ corresponding to counting process $N(t)$ is predictable for a continuous survival time.

Doob-Meyer Decomposition. Assume $N(t)$ is a non-negative submartingale adapted to filtration $\mathcal{F}t$. Then, there exist a right continuous and non-negative predictable process $A(t)$ with finite expectation and

$$M(t) = N(t) - A(t)$$

is a martingale and if further $A(0) = 0$ a.s. then A is determined uniquely.

Predictable variation process. Suppose $M(t)$ is a martingale. Then, by Jensen's inequality $M^2(t)$ is a submartingale. From the Doob-Meyer decomposition, there is a right continuous and non-negative predictable process $V(t)$ with finite expectation such that

$$M^2(t) - V(t)$$

is a martingale. $V(t)$ is called a predictable variation process and denoted as $\langle M, M\rangle(t)$.

Theorem 2 *Suppose that $N_{ij+}(t)$, $j = 1, 2, ..., K, \quad i = 1, ..., n$ are bounded counting processes, $M_{ij+}(t)$, $j = 1, 2, ..., K, \quad i = 1, 2, ..., n$ are the corresponding zero-mean counting process martingales constructed in the form*

$$M_{ij+}(t) = N_{ij+}(t) - A(t) = N_{ij+}(t) - \int_0^t Y_{ij+}(t)\lambda(t)dt$$

where $Y_{ij+}(t)$ is corresponding at risk process. Assume each $M_{ij+}(t)$ satisfies $E(M_{ij+})^2 \leq \infty$ for any t, and $H_{ij+}(t)$ are bounded and predictable processes. Suppose also that the filtration

concerned is right continuous. Let

$$Q_{ij+}(t) = \int_0^t H_{ij+}(s) dM_{ij+}(s) \tag{A.1}$$

Then

$$\langle Q_{ij+}, Q_{i'j+} \rangle(t) = \int_0^t H_{ij+}(s) H_{i'j+}(s) d\langle M_{ij+}, M_{i'j+} \rangle(s).$$

Orthogonal counting process martingales. If $\langle M_1, M_2 \rangle(t) = 0$ almost surely, then $M_1(t)$ and $M_2(t)$ are said to be orthogonal. Additionally, it is worth noting that if $\langle M_1, M_2 \rangle(t) = 0$ almost surely, then the product $M_1(t)M_2(t)$ is also a martingale. Assume $M_{ij+}(t)$, $j = 1, 2, ..., K, \quad i = 1, 2, ..., n$ are orthogonal so are $Q_{ij+}(t)$, $j = 1, 2, ..., K, \quad i = 1, 2, ..., n$, and define $Q_j(t) = \sum_{i=1}^n Q_{ij+}(t)$. Then

$$\text{Var}(Q_j(t)) = \sum_{i=1}^n E(\int_0^t H_{ij+}^2(s) dA_i(s)). \tag{A.2}$$

Furthermore,

$$\text{Cov}(Q_j(t), Q_{j'}(t)) = \sum_{i=1}^n E(\int_0^t H_{ij+}(s) H_{ij'+}(s) dA_i(s)).$$

Next, we present the martingale central limit theorem. First, define the following:

- $U_{in}(t) = \int_0^t H_{in}(s) dM_{in}(s)$,
- $U_n(t) = \sum_{i=1}^n U_{in}(t)$,
- $U_{in,\epsilon}(t) = \int_0^t H_{in}(s) I_{[H_{in}(s) \geq \epsilon]} dM_{in}(s)$, and
- $U_{n,\epsilon}(t) = \sum_{i=1}^n U_{in,\epsilon}(t)$.

Theorem 3 *(Martingale Central Limit Theorem (MCLT))*

Assume the filtration concerned is right-continuous. Also, assume for any n and $i = 1, 2, ..., n$, H_{in} *is left continuous and bounded. Furthermore,*

$$< U_n, U_n > (t) \xrightarrow{p} V(t)$$

where V *is non-random. And, for every* $\epsilon > 0$ *and as* $n \to \infty$

$$< U_{n,\epsilon}, U_{n,\epsilon} > (t) \xrightarrow{p} 0.$$

Then, $n \to \infty$

$$U_n(t) \to U(t)$$

in distribution (weakly), where $U(t)$ *is a zero-mean Gaussian process with independent increments and variance function* $V(t)$.

Using Theorem 3, we can find the limit $V(t)$ by considering

$$< U_n, U_n > (t) = \sum_{i=1}^{n} \int_0^t H_{in}^2(s) dA_{in}(s),$$

and

$$< U_{in,\epsilon}, U_{in,\epsilon} > (t) = \sum_{i=1}^{n} \int_0^t H_{in}^2(s) I_{[H_{in}(s) \geq \epsilon]} dA_{in}(s)$$

A.2 Outline of the proof of Theorem 1

In this section, we present the proof of Theorem 1, focusing solely on the null hypothesis. It is important to emphasize that the proof relies on the assumptions (A1)-(A3) established in

Chapter 4, along with the basic notations introduced therein. First, consider the adjusted statistic

$$\hat{U}_{+j+}^{(adj)} = \frac{1}{n}\sum_{i=1}^{n}\left([\hat{O}_{ij+}^{(1)} - I_{[J_i=j]}(X_i - \bar{X})^T\hat{\beta}_{j+}^{(1)})] - [\hat{O}_{ij+}^{(2)} - I_{[J_i\neq j]}(X_i - \bar{X})^T\hat{\beta}_{j+}^{(2)})]\right). \quad \text{(A.3)}$$

From Lemma and simple algebra, the equation in (A.3) can be written as

$$= \frac{1}{n}\sum_{i=1}^{n}[\hat{O}_{ij+}^{(1)} - I_{[J_i=j]}(X_i - \mu_X)^T\beta_{j+}^{(1)}] + \frac{n_j}{n}\beta_{j+}^{(1)T}(\bar{X} - \mu_X)$$

$$-\frac{1}{n}\sum_{i=1}^{n}[\hat{O}_{ij+}^{(2)} - I_{[J_i\neq j]}(X_i - \mu_X)^T\beta_{j+}^{(2)}] - \frac{(n_j - n)}{n}\beta_{j+}^{(2)T}(\bar{X} - \mu_X) + o_p(\frac{1}{\sqrt{n}})$$

Since $\pi_j \to \frac{n_j}{n}$ a.s., we can substitute $\frac{n_j}{n} = \pi_j + o_p(1)$ and $\frac{(n-n_j)}{n} = (1-\pi_j) + o_p(1)$ and get

$$= \frac{1}{n}\sum_{i=1}^{n}[\hat{O}_{ij+}^{(1)} - I_{[J_i=j]}(X_i - \mu_X)^T\beta_{j+}^{(1)}] + \pi_j\beta_{j+}^{(1)T}(\bar{X} - \mu_X)$$

$$-\frac{1}{n}\sum_{i=1}^{n}[\hat{O}_{ij+}^{(2)} + I_{[J_i\neq j]}(X_i - \mu_X)^T\beta_{j+}^{(2)}] - (1-\pi_j)\beta_{j+}^{(2)T}(\bar{X} - \mu_X) + o_p(\frac{1}{\sqrt{n}})$$

then

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}[\hat{O}_{ij+}^{(1)} - I_{[J_i=j]}(X_i - \mu_X)^T\beta_{j+}^{(1)}]}_{M_1} - \underbrace{\frac{1}{n}\sum_{i=1}^{n}[\hat{O}_{ij+}^{(2)} - I_{[J_i\neq j]}(X_i - \mu_X)^T\beta_{j+}^{(2)}]}_{M_2}$$

$$+\underbrace{(\bar{X} - E(\bar{X}|J_1, J_2, ...J_n))^T(\pi_j\beta_{j+}^{(1)} - (1-\pi_j)\beta_{j+}^{(2)}))}_{M_3} + o_p(\frac{1}{\sqrt{n}}) \quad \text{(A.4)}$$

where we write the M_3 term since $E(\bar{X} \mid J_1, \ldots, J_n) = \mu_X$. We next show that $\sqrt{n}(M_1 - M_2 + M_3)$ is asymptotically normal. We consider the following random vector to be able to use central

limit theorem

$$\sqrt{n}\begin{pmatrix} E_n[\hat{O}^{(1)}_{ij+} - I_{[J_i=j]}(X_i-\mu_X)^T\beta^{(1)}_{j+}] \\ E_n[\hat{O}^{(2)}_{ij+} - I_{[J_i\neq j]}(X_i-\mu_X)^T\beta^{(2)}_{j+}] \\ E_n[X_i-\mu_X)] \end{pmatrix} \tag{A.5}$$

where $E_n[K_i] = \frac{1}{n}\sum_{i=1}^n K_i$. We can see that conditioned on J, every component in (A.5) is an average of independent and identically distributed terms. Similar to the proof of Theorem 1 in Ye et al. (2023), the Lindeberg's Central Limit Theorem justifies that $\sqrt{n}(M_1 - M_2 + M_3)$ is asymptotically normal with mean 0 conditional on J as $n \to \infty$. To calculate the variances, we consider each term in (A.4). Because of the projection we made,

$$\text{var}(\hat{O}^{(1)}_{ij+}) = \text{var}(\hat{O}^{(1)}_{ij+} + I_{[J_i=j]}(X_i-\mu_X)^T\beta^{(1)}_{j+}),$$

and

$$\text{var}(\hat{O}^{(2)}_{ij+}) = \text{var}(\hat{O}^{(2)}_{ij+} + I_{[J_i\neq j]}(X_i-\mu_X)^T\beta^{(2)}_{j+}).$$

Thus

$$\begin{aligned}\text{Var}(\sqrt{n}M_1) &= \frac{1}{n}\sum_{i=1}^n \text{Var}(\hat{O}^{(1)}_{ij+} - I_{[J_i=j]}(X_i-\mu_X)^T\beta^{(1)}_{j+}) \\ &= \pi_j \text{Var}(\hat{O}^{(1)}_{ij+} - I_{[J_i=j]}(X_i-\mu_X)^T\beta^{(1)}_{j+})\end{aligned}$$

and similarly

$$\text{Var}(\sqrt{n}M_2) = (1-\pi_j)\text{Var}(\hat{O}^{(2)}_{ij+} - I_{[J_i\neq j]}(X_i-\mu_X)^T\beta^{(2)}_{j+}).$$

We note that M_1 and M_2 are orthogonal, with covariance 0, because they depend respectively on independent units from different outcome groups. Hence we obtain

$$\text{Var}(\sqrt{n}(M_1 - M_2)) = \pi_j \text{Var}(\hat{O}^{(1)}_{ij+} - I_{[J_i=j]} X_i^T \beta^{(1)}_{j+})) + (1-\pi_j)\text{Var}(\hat{O}^{(2)}_{ij+} - I_{[J_i \neq j]} X_i^T \beta^{(2)}_{j+}).$$

Also,

$$\text{Var}(\sqrt{n}\bar{X}) = E(Var(X_i|J_i)).$$

We know under null hypothesis that the parameters we defined as $\theta^{(1)}_j$ and $\theta^{(2)}_j$ are zero, but we can still insert those to obtain

$$n\text{Cov}(M_1, \bar{X}) = \frac{1}{n}\text{Cov}(I_{[J_i=j]}(\hat{O}^{(1)}_{ij+} - \theta^{(1)}_j - (X_i - \mu_X)^T \beta^{(1)}_{j+})), X_i),$$

and

$$n\text{Cov}(M_2, \bar{X}) = \frac{1}{n}\text{Cov}(I_{[J_i \neq j]}(\hat{O}^{(2)}_{ij+} - \theta^{(2)}_j - (X_i - \mu_X)^T \beta^{(2)}_{j+})), X_i).$$

By using the definition of β_j and θ_j s, we can show

$$E(X_i \left[I_{[J_i=j]} \left(\hat{O}^{(1)}_{ij+} - \theta^{(1)}_j - (X_i - \mu_X)^T \beta^{(1)}_{j+} \right) \right])$$

$$= \pi_j(\text{Cov}(X_i, \hat{O}^{(1)}_{ij+} | J_i = j) - \text{Var}(X_i | J_i = j)\beta^{(1)}_{j+})$$

$$= \pi_j(\text{Cov}(X_i, \hat{O}^{(1)}_{ij+} | J_i = j) - \text{Var}(X_i | J_i = j)(\text{Var}(X_i \mid J_i = j))^{-1}\text{Cov}(X_i,\, \hat{O}^{(1)}_{ij+} \mid J_i = j)) = 0$$

similarly

$$E(X_i \left[I_{[J_i \neq j]} \left(\hat{O}^{(2)}_{ij+} - \theta^{(2)}_j - (X_i - \mu_X)^T \beta^{(2)}_{j+} \right) \right])$$

$$= (1-\pi_j)(\text{Cov}(X_i, \hat{O}^{(2)}_{ij+}|J_i \neq j) - \text{Var}(X_i|J_i \neq j)\beta^{(2)}_{j+})$$

$$= (1-\pi_j)(\text{Cov}(X_i, \hat{O}^{(2)}_{ij+}|J_i \neq j) - \text{Var}(X_i|J_i \neq j)(\text{Var}(X_i|J_i \neq j))^{-1}\text{Cov}(X_i,\, \hat{O}^{(2)}_{ij+} \,|\, J_i \neq j)) = 0.$$

Therefore,

$$n\text{Cov}(M_1, \bar{X}) = n\text{Cov}(M_2, \bar{X}) = 0.$$

Finally, using Slutsky's theorem,

$$\sqrt{n}(M_1 - M_2 + M_3) \xrightarrow{d} \mathcal{N}\Big(0, \pi_j \text{Var}(\hat{O}^{(1)}_{ij+} - I_{[J_i=j]}X_i^T\beta^{(1)}_{j+})) + (1-\pi_j)\text{Var}(\hat{O}^{(2)}_{ij+} - I_{[J_i\neq j]}X_i^T\beta^{(2)}_{j+}))$$
$$+ (\pi_j\beta^{(1)}_{j+} - (1-\pi_j)\beta^{(2)}_{j+}))^T E(\text{Var}(X_i|J_i))(\pi_j\beta^{(1)}_{j+} - (1-\pi_j)\beta^{(2)}_{j+})$$

Finally, by definition of β_j terms, the variance term can be written as

$$\pi_j \text{Var}(O^{(1)}_{ij+} - X_i^T\beta^{(1)}_{j+})) + (1-\pi_j)\text{Var}(O^{(2)}_{ij+} - X_i^T\beta^{(2)}_{j+}))$$

$$+(\pi_j\beta^{(1)}_{j+} - (1-\pi_j)\beta^{(2)}_{j+}))^T E(\text{Var}(X_i|J_i))(\pi_j\beta^{(1)}_{j+} - (1-\pi_j)\beta^{(2)}_{j+})$$

$$= \pi_j \text{Var}(O^{(1)}_{ij+}) + (1-\pi_j)\text{Var}(O^{(2)}_{ij+}) + \pi_j(1-\pi_j)(\beta^{(1)}_{j+} + \beta^{(2)}_{j+})^T\Sigma_X(\beta^{(1)}_{j+} + \beta^{(2)}_{j+})$$

Finally, following the last steps in Ye et al. (2023) proves Theorem 1. □

We now give the justifications for the covariance terms that we need for the construction of K-sample co-variate adjusted test statistic in Equation 4.34. Denote $M_k \equiv \mathcal{M}^{(j)}_k$ (for $k = 1, 2, 3$) for the sums defined in (A.4), corresponding to fixed group j. The following formulas are asymptotic forms, that contain in-probability limits of stochastic integrals plus other limiting

parametric terms, for variances and covariances of terms involving these $\mathcal{M}_k^{(j)}$, $\mathcal{M}_k^{(j')}$ sums, under the null hypothesis. In what follows, always $j \neq j'$.

$$\text{Var}(\sqrt{n}\,\mathcal{M}_1^{(j)}) \;\sim\; \frac{1}{n}\int_0^{\tau_n} \frac{Y_{+j+}\,(Y_{+++}-Y_{+j+})^2}{Y_{+++}^3}\,dN_{+++} \;-\; \pi_j\,\beta_{j+}^{(1)tr}\,\Sigma_X\,\beta_{j+}^{(1)}$$

$$\text{Var}(\sqrt{n}\,\mathcal{M}_2^{(j)}) \;\sim\; \frac{1}{n}\int_0^{\tau_n} \frac{Y_{+j+}^2\,(Y_{+++}-Y_{+j+})}{Y_{+++}^3}\,dN_{+++} \;-\; \pi_j\,\beta_{j+}^{(2)tr}\,\Sigma_X\,\beta_{j+}^{(2)}$$

$$\text{Cov}(\sqrt{n}\,\mathcal{M}_1^{(j)},\,\sqrt{n}\,\mathcal{M}_2^{(j)}) \;=\; \text{Cov}(\sqrt{n}\,\mathcal{M}_1^{(j)},\,\sqrt{n}\,\mathcal{M}_1^{(j')}) \;=\; 0$$

$$\text{Cov}(\sqrt{n}\,\mathcal{M}_1^{(j)},\,\sqrt{n}\,\mathcal{M}_2^{(j')}) \;\sim\; \frac{1}{n}\int_0^{\tau_n} \frac{Y_{+j'+}\,Y_{+j+}\,(Y_{+++}-Y_{+j+})^2}{Y_{+++}^3}\,dN_{+++} - \pi_j\,\beta_{j+}^{(1)tr}\,\Sigma_X\,\beta_{j'+}^{(2)}$$

$$\text{Cov}(\sqrt{n}\,\mathcal{M}_2^{(j)},\,\sqrt{n}\,\mathcal{M}_1^{(j')}) \;\sim\; \frac{1}{n}\int_0^{\tau_n} \frac{Y_{+j+}\,Y_{+j'+}\,(Y_{+++}-Y_{+j'+})^2}{Y_{+++}^3}\,dN_{+++} - \pi_{j'}\,\beta_{j+}^{(2)tr}\,\Sigma_X\,\beta_{j'+}^{(1)}$$

$$\text{Cov}(\sqrt{n}\,\mathcal{M}_2^{(j)},\,\sqrt{n}\,\mathcal{M}_2^{(j')}) \;\sim\; \frac{1}{n}\int_0^{\tau_n} \frac{Y_{+j+}\,Y_{+j'+}\,(Y_{+++}-Y_{+j+}-Y_{+j'+})}{Y_{+++}^3}\,dN_{+++}$$

$$-\,(1-\pi_j-\pi_{j'})\,\beta_{j+}^{(2)tr}\,\Sigma_X\,\beta_{j'+}^{(2)}$$

$$\text{Cov}(\sqrt{n}\,\mathcal{M}_1^{(j)},\,\sqrt{n}\,\mathcal{M}_3^{(j)}) \;\sim\; \text{Cov}(\sqrt{n}\,\mathcal{M}_1^{(j)},\,\sqrt{n}\,\mathcal{M}_3^{(j')}) \;\sim\; 0$$

$$\text{Cov}(\sqrt{n}\,\mathcal{M}_3^{(j)},\,\sqrt{n}\,\mathcal{M}_3^{(k)}) \;\sim\; (\pi_j\,\beta_{j+}^{(1)}-(1-\pi_j)\beta_{j+}^{(2)})^{tr}\,\Sigma_X\,(\pi_k\,\beta_{k+}^{(1)}-(1-\pi_k)\beta_{k+}^{(2)}) \quad , \quad k \;=j,\; j'$$

So, putting all of these terms together,

$$\text{Cov}\Big(\sqrt{n}\,(\mathcal{M}_1^{(j)}-\mathcal{M}_2^{(j)}+\mathcal{M}_s^{(j)}),\,\sqrt{n}\,\sqrt{n}\,(\mathcal{M}_1^{(j')}-\mathcal{M}_2^{(j')}+\mathcal{M}_s^{(j')})\Big) \;\sim$$

$$\frac{-1}{n}\int_0^{\tau_n}\frac{Y_{+j+}\,Y_{+j'+}}{Y_{+++}^3}\,dN_{+++}\;-\;\pi_j\,\beta_{j+}^{(1)tr}\,\Sigma_X\,\beta_{j'+}^{(2)}\;-\;\pi_{j'}\,\beta_{j+}^{(2)tr}\,\Sigma_X\,\beta_{j'+}^{(1)}$$

$$-\,(1-\pi_j-\pi_{j'})\,\beta_{j+}^{(2)tr}\,\Sigma_X\,\beta_{j'+}^{(2)}\;+\;(\pi_j\,\beta_{j+}^{(1)}-(1-\pi_j)\beta_{j+}^{(2)})^{tr}\,\Sigma_X\,(\pi_{j'}\,\beta_{j'+}^{(1)}-(1-\pi_k)\beta_{j'+}^{(2)})$$

$$=\;\frac{-1}{n}\int_0^{\tau_n}\frac{Y_{+j+}\,Y_{+j'+}}{Y_{+++}^3}\,dN_{+++}\;\;+\;\;\pi_j\,\pi_{j'}\,(\beta_j^{(1)}-\beta_j^{(2)})^{tr}\,\Sigma_X\,(\beta_{j'}^{(1)}-\beta_{j'}^{(2)})$$

and this final term with hats on the π, β, Σ_X terms gives $\hat{\Sigma}_{j,j'}$.

Bibliography

S. Anders, P. T. Pyl, and W. Huber. Htseq—a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2015. doi: 10.1093/bioinformatics/btu638.

P. K. Andersen and R. Gill. Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, 10(4):1100–1120, 1982.

Abu-Salih Bilal. Domain-specific knowledge graphs: A survey. *Journal of Network and Computer Applications*, 185:103076, 2021. ISSN 1084-8045. doi: 10.1016/j.jnca.2021.103076.

J. Martin Bland and Douglas G. Altman. The logrank test. *BMJ*, 328(7447):1073, 2004. doi: 10.1136/bmj.328.7447.1073.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi: 10.1023/A:1010933404324.

Leo Breiman and Adele Cutler. Random forests. 2023. URL `https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm`.

C. G. Broyden. The convergence of a class of double-rank minimization algorithms. *Journal of the Institute of Mathematics and its Applications*, 6(1):76–90, 1970.

Julián Candia, Enkhjargal Bayarsaikhan, Mayank Tandon, Anuradha Budhu, Marshonna Forgues, Lkhagva-Ochir Tovuu, Undarmaa Tudev, Justin Lack, Ann Chao, Jigjidsuren CChinburen, and Xin Wei Wang. The genomic landscape of mongolian hepatocellular carcinoma. *Nature Communications*, 11(1):4383, 2020.

G. Castelli, E. Pelosi, and U. Testa. Liver cancer: Molecular characterization, clonal evolution and cancer stem cells. *Cancers (Basel)*, 9(9):127, 2017. doi: 10.3390/cancers9090127.

Jittiporn Chaisaingmongkol, Anuradha Budhu, Huy Dang, Surangkanang Rabibhadana, Boriboon Pupacdi, Sook-Min Kwon, Xin Wei Wang, and TIGER-LC Consortium. Common molecular subtypes among asian hepatocellular carcinoma and cholangiocarcinoma. *Cancer Cell*, 32(1):57–70.e3, 2017. doi: 10.1016/j.ccell.2017.05.009.

Hsin-wen Chang and Ian W. McKeague. Empirical likelihood based tests for stochastic ordering under right censorship. *Electronic Journal of Statistics*, 10(2):2511–2536, 2016. doi: 10.1214/16-EJS1180.

Hyunseok Choo-Wosoba, Paul S. Albert, and Bin Zhu. hseghmm: hidden markov model-based allele-specific copy number alteration analysis accounting for hypersegmentation. *BMC Bioinformatics*, 19:424, 2018. doi: 10.1186/s12859-018-2412-y.

David Collett. *Modelling survival data in medical research*. Texts in statistical science. Chapman and Hall, London, 1994.

Amanda J Craig, Johann von Felden, Teresa Garcia-Lezana, Samantha Sarcognato, and Augusto Villanueva. Tumour evolution in hepatocellular carcinoma. *Nat Rev Gastroenterol Hepatol*, 17:139–152, 2020. doi: 10.1038/s41575-019-0229-4.

Andrew Davis, Ruli Gao, and Nicholas Navin. Tumor evolution: Linear, branching, neutral or punctuated? *Biochim Biophys Acta Rev Cancer*, 1867(2):151–161, 2017. doi: 10.1016/j.bbcan.2017.01.003.

Leandro de Araújo Lima and Kai Wang. Penncnv in whole-genome sequencing data. *BMC Bioinformatics*, 18(Suppl 11):383, 2017. doi: 10.1186/s12859-017-1802-x.

Daniel L. Dexter, Henryk M. Kowalski, Beverly A. Blazar, Zuzana Fligiel, Renee Vogel, and Gloria H. Heppner. Heterogeneity of tumor cells from a single mouse mammary tumor. *Cancer Res*, 38(10):3174–3181, 1978.

L. Dressler, Bortolomeazzi M., Keddar M. R., Misetic H., Sartini G., A. Acha-Sagredo, L. Montorsi, N. Wijewardhane, D. Repana, J. Nulsen, J. Goldman, M. Pollitt, P. Davis, A. Strange, K. Ambrose, and F. D. Ciccarelli. Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the network of cancer genes (ncg) resource. *Genome Biology*, 23(1):35, 2022.

P. Kaitlyn Edelson, Lorraine Dugoff, and Bryann Bromley. Chapter 11 - genetic evaluation of fetal sonographic abnormalities. In Mary E. Norton, Jeffrey A. Kuller, and Lorraine Dugoff, editors, *Perinatal Genetics*, pages 105–124. Elsevier, 2019. ISBN 9780323530941. URL `https://doi.org/10.1016/B978-0-323-53094-1.00011-4`.

Camilla Engblom, Christina Pfirschke, and Mikael J. Pittet. The role of myeloid cells in cancer therapies. *Nature Reviews Cancer*, 16(7):447–462, 2016.

Francesca Favero, Tejal Joshi, Andrea M. Marquard, Nicolai Juul Birkbak, Marcin Krzystanek, Qiyuan Li, and Aron C. Eklund. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology*, 26(1):64–70, 2015. doi: 10.1093/annonc/mdu479.

Thomas Fleming and David Harrington. *Counting Processes and Survival Analysis*. Wiley, New York, 1991.

G. H. Freeman and J. H. Halton. Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika*, 38(1-2):141–149, 1951.

Jane Fridlyand, Antoine M. Snijders, Daniel Pinkel, Donna G. Albertson, and Ajay N. Jain. Hidden markov models approach to the analysis of array cgh data. *Journal of Multivariate Analysis*, 90(1):132–153, 2004. ISSN 0047-259X. doi: 10.1016/j.jmva.2004.02.008. URL https://www.sciencedirect.com/science/article/pii/S0047259X04000260.

J. Friemel, M. Rechsteiner, L. Frick, F. Böhm, K. Struckmann, M. Egger, H. Moch, M. Heikenwalder, and A. Weber. Intratumor heterogeneity in hepatocellular carcinoma. *Clinical Cancer Research*, 21(8):1951–1961, 2015.

E. A. Gehan. A generalized wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika*, 52:203–223, 1965.

Christian Gilissen, Alexander Hoischen, Han G. Brunner, and Joris A. Veltman. Disease gene identification strategies for exome sequencing. *Eur J Hum Genet*, 20(5):490–497, 2012. doi: 10.1038/ejhg.2011.258. Epub 2012 Jan 18.

Robert L. Grossman, Allison P. Heath, Vincent Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, and Louis M. Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016.

John A. Hartigan and Manchek A. Wong. Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, 28(1):100–108, 1979.

G. H. Heppner. Tumor heterogeneity. *Cancer Res*, 44(6):2259–2265, 1984.

Richard R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201, 1983.

Paul Jaccard. The distribution of the flora in the alpine zone.1. *New Phytologist*, 11(2):37–50, 1912. ISSN 0028-646X. URL https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(457–481), 1958. doi: 10.1080/01621459.1958.10501452.

Leonard Kaufman and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 1990.

Daehwan Kim, Ben Langmead, and Steven L. Salzberg. Hisat: a fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360, 2015. doi: 10.1038/nmeth.3317.

Fanhui Kong and Eric Slud. Robust covariate-adjusted logrank tests. *Biometrika*, 84(4):847–862, 1997. doi: 10.1093/biomet/84.4.847.

Kimberly R. Kukurba and Stephen B. Montgomery. Rna sequencing and analysis. *Cold Spring Harb Protoc*, 2015(11):951–969, 2015. doi: 10.1101/pdb.top084970.

Thomas LaFramboise. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research*, 37(13):4181–4193, 2009. doi: 10.1093/nar/gkp552.

Depeng Li. Statistical methods for rna sequencing data analysis. In Holger Husi, editor, *Computational Biology*, chapter 6. Codon Publications, 2019. doi: 10.15586/computationalbiology.2019.ch6. Available from: `https://www.ncbi.nlm.nih.gov/books/NBK550334/`.

Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. doi: 10.1093/bioinformatics/btp324.

Y. Liao, G. K. Smyth, and W. Shi. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, 41(10):e108, 2014. doi: 10.1093/nar/gkt214.

Chia-Wei Lin, Adam C. Naj, and Li-San Wang. Analyzing copy number variation using snp array data: protocols for calling cnv and association tests. *Current Protocols in Human Genetics*, 79:1.27.1–1.27.15, 2013. doi: 10.1002/0471142905.hg0127s79.

Patrick Little, Daowen Y. Lin, and Wei Sun. Associating somatic mutations to clinical outcomes: a pan-cancer study of survival time. *Genome Medicine*, 11(1):37, 2019.

Liangcai Ma, Zhe Chen, Marco Erdmann, Zhilong Zhang, Di Wu, Hongyu Cao, and Hongyang Wang. Multiregional single-cell dissection of tumor and immune cells reveals stable lock-and-key features in liver cancer. *Nature Communications*, 13(1):7533, 2022.

Kumar P. Mainali, Eric Slud, Michael C. Singer, and William F. Fagan. A better index for analysis of co-occurrence and similarity. *Science Advances*, 8(4):eabj9204, 2022.

N. Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50:163–170, 1966.

Nicholas McGranahan and Charles Swanton. Clonal heterogeneity and tumor evolution: Past, present, and the future. *Cell*, 168(4):613–628, 2017. doi: 10.1016/j.cell.2017.01.018.

Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, and DePristo Mark A. Kernytsky, Andrew. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research*, 20(9):1297–1303, 2010. doi: 10.1101/gr.107524.110.

A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, and A. A. Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12(5):453–457, 2015. doi: 10.1038/nmeth.3337.

Gro Nilsen, Knut Liestøl, Peter Van Loo, Hans Kristian Moen Vollan, Morten B. Eide, Oscar M. Rueda, and Ole Christian Lingjærde. Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics*, 13:591, 2012. doi: 10.1186/1471-2164-13-591.

Peter C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

Adam B. Olshen, E. S. Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572, 2004. doi: 10.1093/biostatistics/kxh008.

C. Ormond, N.M. Ryan, A. Corvin, and E.A. Heron. Converting single nucleotide variants between genome builds: from cautionary tale to solution. *Briefings in Bioinformatics*, 22 (5):bbab069, 2021. doi: 10.1093/bib/bbab069.

M. Overgaard and S. Hansen. On the assumption of independent right censoring. *Scandinavian Journal of Statistics*, 48:1234–1255, 2019. doi: 10.1111/sjos.12487.

Daniel A. Peiffer, Jennifer M. Le, Frank J. Steemers, Wenwei Chang, Todd Jenniges, and Gunderson Kevin L. Garcia, Francisco. High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. *Genome Research*, 16(9):1136–1148, 2006. doi: 10.1101/gr.5402306.

R. Peto and J. Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A (General)*, 135(2):185–206, 1972. doi: 10.2307/2344317.

Richard Peto, Malcolm C. Pike, Peter Armitage, Norman E. Breslow, David R. Cox, Simon V. Howard, Peter Smith, N. G.Mantel, K. McPherson, J. Peto, and P. G. Smith. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. i. introduction and design. *Br J Cancer*, 34(6):585–612, 1976. doi: 10.1038/bjc.1976.220.

L.C.G. Rogers and David Williams. *Diffusions, Markov processes, and martingales*, volume 1 of *Wiley series in probability and mathematical statistics*. John Wiley and Sons, Chichester, New York, 1994.

A. Roth, Khattra J., D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté, and S. P. Shah. Pyclone: statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398, 2014.

P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Brian Ruffell and Lisa M. Coussens. Macrophages and therapeutic resistance in cancer. *Cancer Cell*, 27(4):462–472, 2015.

M. Schmid, M. Oudot, and J. P. Vert. Estimating the parameters of mixtures of multidimensional Gaussian distributions. Technical Report 95(34), INRIA, 1992.

Ronglai Shen and Venkatraman E. Seshan. Facets: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput dna sequencing. *Nucleic Acids Research*, 44 (16):e131, 2016. doi: 10.1093/nar/gkw520.

C. B. Steen, C. L. Liu, A. A. Alizadeh, and A. M. Newman. Profiling cell type abundance and expression in bulk tissues with cibersortx. *Methods in Molecular Biology*, 2117:135–157, 2020.

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

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

The Cancer Genome Atlas Research Network. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell*, 169(7):1327–1341.e23, 2017.

Terry M. Therneau. *A Package for Survival Analysis in R*, 2023. URL `https://CRAN.R-project.org/package=survival`. R package version 3.5-5.

Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000. ISBN 0-387-98784-3.

R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63 (2):411–423, 2001.

Peter Van Loo, Silje H. Nordgard, Ole Christian Lingjærde, Hege G. Russnes, Inga H. Rye, Wenyu Sun, and Vessela N. Kristensen. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915, 2010. doi: 10.1073/pnas.1009843107. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1009843107`.

Roberto Vendramin, Kevin Litchfield, and Charles Swanton. Cancer evolution: Darwin and beyond. *Embo j*, 40(18):e108389, 2021.

Kai Wang, Mingyao Li, Dexter Hadley, Ryan Liu, Joseph Glessner, Struan F. A. Grant, and Maja Bucan. Penncnv: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. *Genome Research*, 17(11): 1665–1674, 2007. doi: 10.1101/gr.6861907.

Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, 2009. doi: 10.1038/nrg2484.

Andrew Warr, Christel Robert, David Hume, Alan Archibald, Numan Deeb, and Mick Watson. Exome sequencing: Current and future perspectives. *G3 (Bethesda)*, 5(8):1543–1550, 2015. doi: 10.1534/g3.115.018564.

Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002. URL https://proceedings.neurips.cc/paper_files/paper/2002/file/c3e4035af2a1cde9f21e1ae1951ac80b-Paper.pdf.

Tianxi Ye, Jun Shao, and Yanping Yi. Covariate-adjusted log-rank test: Guaranteed efficiency gain and universal applicability. *arXiv preprint arXiv:2201.11948v2 [stat.ME]*, Jan 2023.

M. Zhang, J. L. Luo, Q. Sun, J. Harber, A. G. Dawson, A. Nakas, S. Busacca, A. J. Sharkey, D. Waller, M. T. Sheaff, C. Richards, P. Wells-Jordan, A. Gaba, C. Poile, E. Y. Baitei, A. Bzura, J. Dzialo, M. Jama, J. Le Quesne, A. Bajaj, and D. A. Fennell. Clonal architecture in mesothelioma is prognostic and shapes the tumor microenvironment. *Nature Communications*, 12(1):1751, 2021.

Z. Zhang, X. Zhou, H. Shen, D. Wang, Y. Wang, K. Phan, and Q. Gao. Landscape of infiltrating b cells and their clinical significance in human hepatocellular carcinoma. *Oncoimmunology*, 8 (4):e1571388, 2019.

H. Zhao, Z. Sun, J. Wang, H. Huang, J. P. Kocher, and L. Wang. Crossmap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, 30(7):1006–1007, 2014. doi: 10.1093/bioinformatics/btt730.

Li Zhao, Hongjian Liu, Xiaofei Yuan, Kaitai Gao, and Jun Duan. Comparative study of whole exome sequencing-based copy number variation detection tools. *BMC Bioinformatics*, 21(1): 97, 2020. doi: 10.1186/s12859-020-3421-1.

X. Zhu, S. Li, B. Xu, and H. Luo. Cancer evolution: A means by which tumors evade treatment. *Biomedicine Pharmacotherapy*, 133:111016, 2021.